Low Rates of Expression Profile Divergence in Highly Expressed Genes and Tissue-Specific Genes During Mammalian Evolution

Ben-Yang Liao and Jianzhi Zhang

Department of Ecology and Evolutionary Biology, University of Michigan

Evolutionary rates provide important information about the pattern and mechanism of evolution. Although the rate of gene sequence evolution has been well studied, the rate of gene expression evolution is poorly understood. In particular, it is unclear whether the gene expression level and tissue specificity influence the divergence of expression profiles between orthologous genes. Here we address this question using a microarray data set comprising the expression signals of 10,607 pairs of orthologous human and mouse genes from over 60 tissues per species. We show that the level of gene expression and the degree of tissue specificity are generally conserved between the human and mouse orthologs. The rate of gene expression profile change during evolution is negatively correlated with the level of gene expression, measured by either the average or the highest level among all tissues examined. This is analogous to the observation that the rate of gene (or protein) sequence evolution is negatively correlated with the gene expression level. The impacts of the degree of tissue specificity on the evolutionary rate of gene sequence and that of expression profile, however, are opposite. Highly tissue-specific genes tend to evolve rapidly at the gene sequence level but slowly at the expression profile level. Thus, different forces and selective constraints must underlie the evolution of gene sequence and that of gene expression.

Introduction

It has been proposed that evolutionary changes of morphology and development are more often due to alterations of gene expressions than protein sequences (King and Wilson 1975; Carroll 2005). However, compared to our knowledge of gene and protein sequence evolution (Li 1997; Nei and Kumar 2000), genome-wide patterns of gene expression evolution (Cavalieri, Townsend, and Hartl 2000; Enard et al. 2002; Oleksiak, Churchill, and Crawford 2002; Ranz et al. 2003; Rifkin, Kim, and White 2003) are poorly understood, except for the divergences of duplicate genes (Gu et al. 2002; Makova and Li 2003; Gu et al. 2004; Huminiecki and Wolfe 2004; Gu, Zhang, and Huang 2005; He and Zhang 2005). The advancement of high-throughput technologies for characterizing the expressions of thousands of genes simultaneously and the subsequent availability of microarray expression data from multiple species open the door for searching for general principles governing gene expression evolution. Two recent studies suggested that expression evolution is largely neutral, with little influences of either positive or purifying selection (Khaitovich et al. 2004; Yanai, Graur, and Ophir 2004). However, subsequent experimental studies and computational analysis using microarray-based expression data suggested that the expression evolution of most genes is subject to purifying selection (Denver et al. 2005; Jordan, Marino-Ramirez, and Koonin 2005; Rifkin et al. 2005; Liao and Zhang 2006). For example, Liao and Zhang (2006) estimated that 84% of mammalian genes have significantly lower expression divergence than expected under complete neutrality. These findings raise the question about the determinants of the level of purifying selection on gene expression.

Evolutionary changes of gene expression can be studied from two aspects: (1) changes of gene expression level in a given tissue or under a certain condition and (2) changes of gene expression profile across spatial, temporal, or environmental dimensions. The first aspect has been studied more than the second (e.g., Ranz et al. 2003; Khaitovich et al. 2004). Therefore, we focus on the second aspect in this work. Specifically, we examine expression profile evolution of mammalian genes across tissues by comparing human and mouse orthologs. Pearson's correlation coefficient (r) is used to measure the expression profile similarity between a pair of orthologous genes. Because all human-mouse orthologs have diverged for the same amount of time, one can use r to compare the relative rates of expression profile evolution among genes. That is, higher r indicates a lower rate of evolution, whereas lower r indicates a higher rate of evolution.

Here we consider two potential determinants of the rate of gene expression profile evolution: expression level and tissue specificity. These two factors were previously shown to be major determinants of the rate of gene (or protein) sequence evolution (Hastings 1996; Duret and Mouchiroud 2000; Pal, Papp, and Hurst 2001; Subramanian and Kumar 2004; Zhang and Li 2004; Zhang and He 2005), and our analysis would answer whether gene sequence evolution and expression profile evolution are governed by the same rules. Furthermore, a recent study showed that the expression divergence between a pair of human-mouse orthologs is negatively correlated with the number of tissues in which the gene is expressed (Yang, Su, and Li 2005). This finding is puzzling because highly specific tissue expression of a gene indicates that the gene performs a tissue-specific function (e.g., chemoreception or immunity), and it would be unlikely for such a highly specialized gene to perform functions useful to other tissues in a different species. Here we analyze the Gene Atlas V2 microarray data set (Su et al. 2004), which includes the expression information of 10,607 human and mouse orthologous genes in over 60 tissues. Our analysis indicates that the evolutionary rate of gene expression profile is negatively correlated with the level of expression and the degree of tissue specificity.

Materials and Methods

Gene Expression Data

We used the human and mouse gene expression information from the Gene Atlas V2 data set (http://symatlas.

Key words: evolutionary rate, expression profile, expression level, tissue specificity, mammals.

E-mail: jianzhi@umich.edu.

Mol. Biol. Evol. 23(6):1119–1128. 2006
doi:10.1093/molbev/msj119
Advance Access publication March 6, 2006

gnf.org/), which contains the expression data obtained by hybridization of RNAs from 73 human nonpathogenic tissues and 61 mouse tissues onto the Affymetrix microarray chips (human, U133A/GNF1H; mouse, GNF1M) designed according to the annotated human and mouse genome sequences (Su et al. 2004). A gene is represented on a chip by at least one probe set, each of which comprises either 11 (in human arrays) or 10 (in mouse arrays) pairs of probes that overlap in their nucleotide sequences. To assign the probe sets to the current annotated version of Ensembl human and mouse genes, we aligned sequences of each probe set to the Ensembl cDNA sequences (human, Homo_sapiens.NCBI35.feb.cdna.fa; mouse, Mus_musculus.NCBIM33.feb.cdna.fa; http://www.ensembl.org/info/data/download.html/) using BlastN (http://www.ncbi.nlm.nih.gov/blast/) and kept those probe sets in which all 11 human or 10 mouse matching probes perfectly matched to the same Ensembl gene. For further analysis, 25,368 probe sets (75.3%) in the human chip corresponding to 16,456 genes and 18,005 probe sets (49.8%) in the mouse chip corresponding to 15,835 genes were retained. The expression level detected by each probe set was obtained as the signal intensity (S) computed from either MAS 5.0 algorithm (MAS5) (Hubbell, Liu, and Mei 2002) or GC content–adjusted robust multiarray algorithm (GC-RMA) (Wu et al. 2004). The Gene Atlas V2 data set derived from GC-RMA algorithm was downloaded from GNF Genome Informatics Applications & Data sets (http://wombat.gnf.org). The S values were averaged among replicates. Because the results from MAS5 and GC-RMA are similar, we present the findings obtained from MAS5 unless otherwise noted.

Tissue Specificity of Gene Expression

We used tissue specificity index τ (Yanai et al. 2005) to measure the tissue specificity of a human or mouse gene. The τ of human gene i is defined by

$$\tau_H = \frac{\sum_{j=1}^{n_H}\left(1 - \left[\frac{\log_2 S_H(i,j)}{\log_2 S_H(i,\max)}\right]\right)}{n_H - 1}, \quad (1)$$

where n_H is the number of human tissues examined and $S_H(i, \max)$ is the highest expression signal of gene i across the n_H tissues. To minimize the influence of noise from low intensities, we arbitrarily let $S_H(i, j)$ be 100 if it is lower than 100. Note that this strategy of reducing the effect of noise is used only in computing τ. When a gene has more than one probe set on the chip, we compute τ by averaging the τ values derived from the different probe sets. The τ value ranges from 0 to 1, with higher values indicating higher variations in expressional level across tissues or higher tissue specificities. If a gene has expression in only one tissue, τ approaches 1. In contrast, if a gene is equally expressed in all tissues, $\tau = 0$.

Human-Mouse Orthologs

The homology information of human and mouse genes was obtained from Ensembl EnsMart (http://www.ensembl.org/Multi/martview) (Kasprzyk et al. 2004). There are several annotated homology relationships between human and mouse genes by Ensembl. We only considered those pairs of genes annotated as UBRH (Unique Best Reciprocal Hit, meaning that they were unique reciprocal best hits in all-against-all BlastZ searches) to be orthologous. We found that 10,607 pairs of human-mouse orthologs have expression data. Affymetrix probes with name suffixes _x_at and _s_at were thought to be prone to cross-hybridization, compared to other probes (Affymetrix Technical Support, Data Analysis Fundamentals, Appendix B; http://www.affymetrix.com/support/downloads/manuals/), and have been considered "suboptimal" (Yang, Su, and Li 2005). But our recent analysis showed that the quality of these probes is not worse than other probes (Liao and Zhang 2006). We therefore considered all probe sets equally.

The number of synonymous substitutions per synonymous site (d_S) and the number of nonsynonymous substitutions per nonsynonymous site (d_N) between human and mouse orthologs were retrieved from Ensembl EnsMart. In this database, d_S and d_N were estimated by the maximum likelihood method using the PAML package (Yang 1997).

Expression Profile Similarity Between Orthologous Genes

To measure the similarity in expression profile between human and mouse orthologs, we analyzed 26 common tissues of the two species included in the data set (Su et al. 2004). These 26 tissues are adipocyte, adrenal gland, amygdala, bone marrow, cerebellum, dorsal root ganglion, heart, hypothalamus, kidney, liver, lung, lymph node, ovary, pancreas, pituitary, placenta, prostate, skeletal muscle, spinal cord, testis, thymus, thyroid, tongue, trachea, trigeminal ganglion, and uterus. Mouse lower spinal cord was used as the homologous tissue of human spinal cord. We measured the expression profile similarity between a pair of orthologous genes by Pearson's correlation coefficient r, defined as

$$r = \frac{\sum_{j=1}^{n}[S_H(i,j)S_M(i,j)] - \left[\sum_{j=1}^{n} S_H(i,j)\right]\left[\sum_{j=1}^{n} S_M(i,j)\right]\Big/ n}{\sqrt{\sum_{j=1}^{n}[S_H(i,j)]^2 - \left[\sum_{j=1}^{n} S_H(i,j)\right]^2\Big/ n}\sqrt{\sum_{j=1}^{n}[S_M(i,j)]^2 - \left[\sum_{j=1}^{n} S_M(i,j)\right]^2\Big/ n}}. \quad (2)$$

Here, $n = 26$ is the number of common tissues considered, H indicates human, and M indicates mouse. $S_H(i, j)$ and $S_M(i, j)$ are the expression signal intensities of gene i in human tissue j and mouse tissue j, respectively. A high r indicates a high similarity in expression profile between the orthologs and a low rate of expression profile evolution. Note that in our previous study (Liao and Zhang 2006), the relative abundance of mRNA across tissues (the signal of one tissue

relative to the total signal of all tissues) was used to compute r. In fact, using either relative abundance or S gives exactly the same r value. To compare our results with those of Yang, Su, and Li (2005), we also used the expression conservation index (ECI) that they developed. The ECI between a pair of human-mouse orthologs is

$$\mathrm{ECI} = \frac{N_{\mathrm{HM}} + 0.5}{(N_{\mathrm{H}} + N_{\mathrm{M}})/2 + 0.5}, \quad (3)$$

where N_{H} and N_{M} are the numbers of human and mouse tissues in which the gene is expressed, respectively, and N_{HM} is the number of tissues in which the gene is expressed in both species. According to Yang, Su, and Li (2005), a gene is considered to be expressed in a tissue if $S \geq 200$ for the tissue. ECI varies from 0 to 1, with higher values indicating higher similarity between expression profiles. When a gene is represented by more than one probe set on a microarray chip, r and ECI are computed by averaging the values obtained from all possible combinations of a human probe set and a mouse probe set of the gene.

Results and Discussion

Choice of Parameters Used in This Study

The transcriptome data analyzed in the present study were obtained from oligonucleotide microarray experiments. It is important to consider properties of microarray data when quantifying tissue specificity of a gene or expression profile similarity between a pair of orthologs.

Tissue specificity of gene expression measures the degree of differential expression across tissues. It is expected that a gene with higher tissue specificity tends to have lower expression breadth (B), which is the proportion of tissues in which the gene is expressed. In microarray data analysis, the number of tissues (N) in which a gene is expressed is usually determined by an arbitrary cutoff of the signal intensity S (Su et al. 2002; Vinogradov 2004; Yang, Su, and Li 2005). Similar definitions of tissue specificity have also been used in studies based on serial analysis of gene expression or expression sequence tag data (Duret and Mouchiroud 2000; Ponger, Duret, and Mouchiroud 2001; Lercher, Urrutia, and Hurst 2002; Subramanian and Kumar 2004). However, there are several problems with the approach of applying a cutoff in defining whether a gene is expressed in a tissue. First, the number of mRNA molecules of a gene in a given tissue is a continuous figure; expression should not be characterized as absent or present. Second, the expression level required for a gene to be functional presumably varies substantively among genes; it is unreasonable to use a single cutoff for all genes in all tissues. Third, expression breadth actually measures the tissue restriction of expression but ignores quantitative variations in expression among many tissues (Schug et al. 2005). Fourth, the S value in microarray data is not only determined by the quantity of the target mRNA but also by the probe-target affinity and the algorithm of raw-data processing. In other words, two genes with the same S values do not necessarily have the same mRNA concentration. Although a recent study (Khaitovich et al. 2005) used the Affymetrix detection P value instead of the cutoff value of S to determine the expression status of a gene in a given tissue, several of the above problems cannot be avoided. Because of these problems with the cutoff-based expression breadth (B), we use tissue specificity index (τ) to measure tissue specificity. Use of the parameter τ can avoid the aforementioned problems.

Another potential measure of tissue specificity is the coefficient of variation (CV) of expression across tissues. CV is defined as the standard deviation (SD) of a random variable divided by its mean. The CV value for human gene i can be computed by the SD of $\log_2 S_{\mathrm{H}}(i,j)$ among the 73 human tissues considered divided by the average $\log_2 S_{\mathrm{H}}(i,j)$of the 73 tissues. A high CV indicates a great variation in gene expression among tissues, implying tissue specificity. Because τ and CV are highly correlated (Spearman's rank correlation coefficient = 0.693, $P < 10^{-300}$; Pearson's correlation coefficient = 0.690, $P < 10^{-300}$; see Fig. S1, Supplementary Material online), we will only use τ in this study.

We use Pearson's correlation r to measure the expression profile similarity (conservation) between a pair of orthologous genes. It was claimed by Yang, Su, and Li (2005) that compared to r, ECI is a more appropriate measure. However, our results suggest that r is better than ECI in quantifying expression profile similarity (see below). For instance, unlike ECI, using r avoids the use of cutoff-based method in defining N.

Conservation of Gene Expression Level and Tissue Specificity During Evolution

We examine whether the level of gene expression and the degree of tissue specificity are similar between human and mouse orthologous genes. If gene expression evolution is not selectively constrained, as suggested earlier (Khaitovich et al. 2004; Yanai, Graur, and Ophir 2004), no such similarity is expected (Jordan, Marino-Ramirez, and Koonin 2005) because of the long divergence time between the two species (Springer et al. 2003; Murphy, Pevzner, and O'Brien 2004) and the rapid pace with which gene expression can change during evolution (Gu et al. 2002). However, we found a strong positive correlation in both mean expression level (fig. 1*a*; Spearman's rank correlation coefficient = 0.392, $P < 10^{-300}$) and tissue specificity (fig. 1*b*; Spearman's rank correlation coefficient = 0.296, $P < 10^{-212}$) between human and mouse orthologs. Note that the mean expression levels are calculated from averaging the S values of 73 normal human tissues or 61 mouse tissues. Similar results were obtained when only the 26 common tissues between humans and mice were considered (Spearman's rank correlation coefficient = 0.384, $P < 10^{-300}$, for mean expression level; Spearman's rank correlation coefficient = 0.335, $P < 10^{-276}$, for tissue specificity). It is interesting to note that although the type of microarray data we analyzed were reported to be noisy (Hill et al. 2001; Irizarry et al. 2003) and probe sets of orthologous genes often have different hybridization behaviors (Liao and Zhang 2006), significant similarities in expression level and tissue specificity are still apparent between human and mouse orthologs, strongly suggesting the evolutionary conservation of gene expression. Our result regarding the conservation of gene expression level is consistent with that of Jordan, Marino-Ramirez, and Koonin (2005).

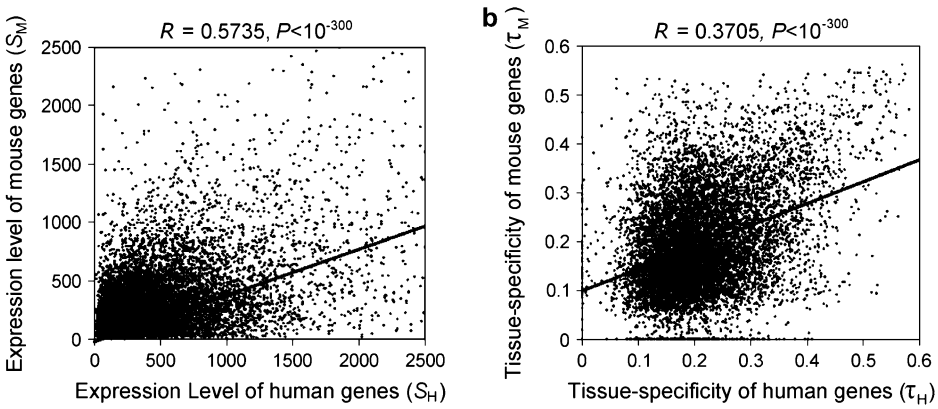

FIG. 1.—Similarity between human-mouse orthologs in (*a*) mean expression level and (*b*) tissue specificity. Spearman's rank correlation coefficient = 0.392 ($P < 10^{-300}$) for (*a*) and 0.296 ($P < 10^{-212}$) for (*b*). In addition, the linear regression and Pearson's correlation coefficient (R) are presented for each panel. The data included 10,607 human-mouse orthologs. The mean expression levels (S_H or S_M) and tissue specificity (τ_H or τ_M) of the human and mouse genes are calculated from 73 human and 61 mouse normal tissues, respectively.

Previous studies showed that gene expression level and expression breadth are strongly and positively correlated (Lercher, Urrutia, and Hurst 2002; Vinogradov 2004). This is not unexpected as expression breadth is determined by the expression signal cutoff used. However, in the present study, virtually no correlation is found between expression level and tissue specificity τ. For example, in humans, Spearman's rank correlation coefficient between τ and mean S is -0.007 ($P = 0.481$). Because the correlations we report in the following two sections are much higher and very significant, it is appropriate to assume that τ and S are uncorrelated.

Highly Expressed Genes Have Low Rates of Expression Profile Evolution

The phenomenon that highly expressed genes have lower substitution rates than lowly expressed genes in coding sequences has been reported in bacteria (Rocha and Danchin 2004), unicellular eukaryotes (Pal, Papp, and Hurst 2001; Wall et al. 2005; Zhang and He 2005), and multicellular eukaryotes (Subramanian and Kumar 2004; Jordan, Marino-Ramirez, and Koonin 2005). This is also true in our data set. For example, the average expression level of human genes (S_H) and the nonsynonymous nucleotide distance d_N between human and mouse orthologs are negatively correlated (Spearman's rank correlation coefficient = -0.160, $P < 10^{-58}$). We also found a weak negative correlation between S_H and the synonymous nucleotide distance d_S (Spearman's rank correlation coefficient = -0.099, $P < 10^{-23}$). S_H and d_N/d_S are also negatively correlated (Spearman's rank correlation coefficient = -0.139, $P < 10^{-44}$). These results confirm that genes of high expression are more selectively constrained in the coding sequence than genes of low expression. Below, we examine whether highly expressed genes are also more constrained in their expression profile evolution.

Our analysis of 10,607 human-mouse orthologs shows that highly expressed genes have more similar expression profiles between species than lowly expressed genes (fig. 2 for the binned data). This is true regardless of whether the expression level is measured by the average S over all tissues (fig. 2*a*, human; fig. 2*b*, mouse) or by the maximum S (fig. 2*c*, human; fig. 2*d*, mouse) among 73 human or 61 mouse tissues examined. For the unbinned original data, the positive correlation between profile similarity and expression level is also strong (rank correlation coefficient: 0.17–0.37) (fig. 2 legend). Because the expression profile similarities are derived from the 26 tissues common to the humans and mice, we also conducted the correlation analysis using average S and maximum S computed from the 26 common tissues. The results obtained (Fig. S2, Supplementary Material online) are similar to those presented in fig. 2. Furthermore, we used the GC-RMA expression data set and obtained similar results (Fig. S3, Supplementary Material online).

It is possible that the positive correlation between gene expression level and expression profile similarity is due to the relatively strong background noise at low expression levels, which would reduce the expression profile similarity more for lowly expressed genes. If our result is mainly due to such a factor, the correlation between the expression level and profile similarity should be much weaker in the subset of genes with high expressions. We examined genes with average $S \geq 800$, a much greater value than that commonly thought to be significant ($S = 200$, Su et al. 2002). We found that highly expressed genes ($S \geq 800$) still show the same trend (fig. 2*a* and *b*), suggesting that our observation is not due to the background noise. Our results thus suggest that highly expressed genes are exposed to stronger purifying selection in both coding sequence evolution and expression profile evolution than lowly expressed genes.

Tissue-Specific Genes Have Low Rates of Expression Profile Evolution

Previous studies showed that broadly expressed genes such as housekeeping genes have lower substitution rates in their coding sequences than narrowly expressed genes (Hastings 1996; Duret and Mouchiroud 2000; Winter, Goodstadt, and Ponting 2004; Zhang and Li 2004). It is expected that the same trend exists between tissue specificity τ and the rate of coding sequence evolution. Indeed, we found weak positive correlations between τ_H and d_N

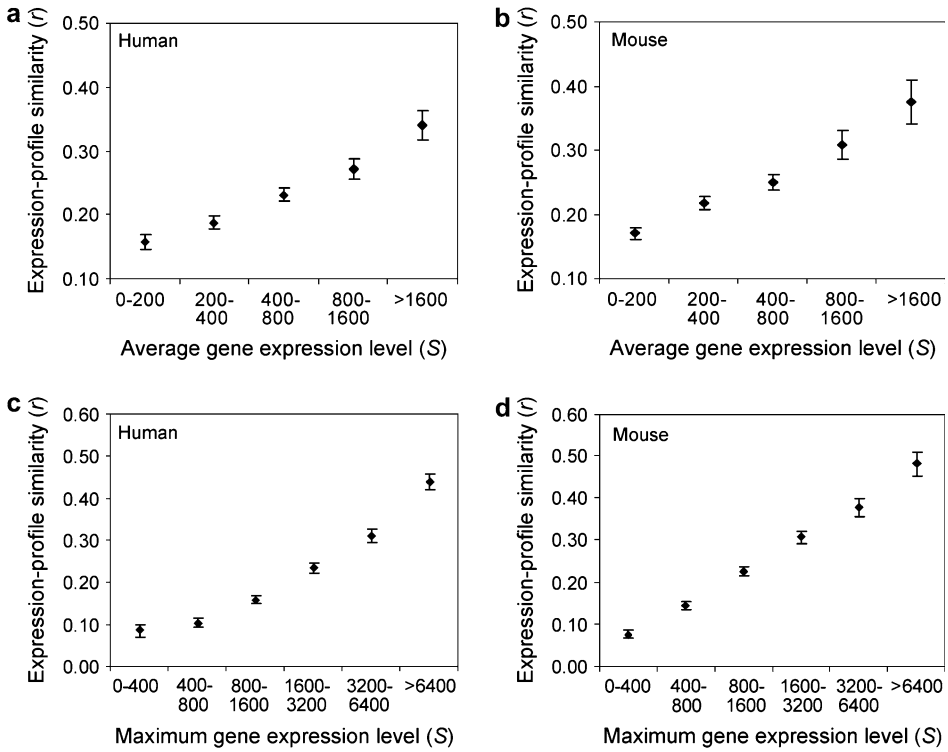

FIG. 2.—Highly expressed genes have higher expression profile similarity between human-mouse orthologs than lowly expressed genes (MAS5 data set). The expression level is measured by either the mean expression level or the maximum expression level across all tissues (i.e., 73 human normal tissues or 61 mouse tissues). The error bar shows 95% confidence interval of the mean, estimated by 10,000 bootstrap replications for each bin. The data include 10,607 human-mouse orthologs. We measured the correlations using the original unbinned data. Spearman's rank correlation coefficient is (*a*) 0.172 ($P < 10^{-71}$), (*b*) 0.176 ($P < 10^{-74}$), (*c*) 0.333 ($P < 10^{-272}$), and (*d*) 0.365 ($P < 10^{-300}$). The numbers of gene pairs used in each bin are (*a*) 0–200: 2,517; 200–400: 2,781; 400–800: 3,093; 800–1,600: 1,576; >1,600: 640; (*b*) 0–200: 4,377; 200–400: 3,132; 400–800: 2,064; 800–1,600: 768; >1,600: 266; (*c*) 0–400: 909; 400–800: 1,900; 800–1,600: 2,743; 1,600–3,200: 2,302; 3,200–6,400: 1,472; >6,400: 1,281; (*d*) 0–400: 2,439; 400–800: 2,507; 800–1,600: 2,402; 1,600–3,200: 1,619; 3,200–6,400: 961; >6,400: 679.

(Spearman's rank correlation coefficient = 0.089, $P < 10^{-18}$), d_S (Spearman's rank correlation coefficient = 0.114, $P < 10^{-24}$), and d_N/d_S (Spearman's rank correlation coefficient = 0.060, $P < 10^{-9}$). Next, we examined the relationship between tissue specificity and the rate of expression profile divergence. We found that genes with higher τ tend to show higher expression profile similarity (r) between human-mouse orthologs (see fig. 3 for the binned

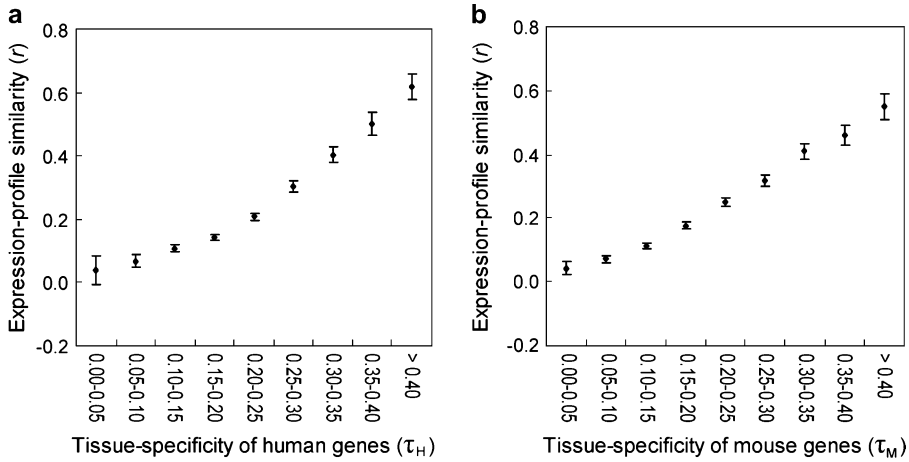

FIG. 3.—Greater expression profile similarities between human-mouse orthologs for genes of high tissue specificity than genes of low tissue specificity (MAS5 data set). Tissue specificity is measured using all tissues (i.e., 73 human normal tissues or 61 mouse tissues). The error bar shows 95% confidence interval of the mean, estimated by 10,000 bootstrap replications for each bin. The data include 10,607 human-mouse orthologs. We measured the correlations using the original unbinned data. Spearman's rank correlation coefficient is (*a*) 0.340 ($P < 10^{-285}$) and (*b*) 0.377 ($P < 10^{-300}$). The numbers of genes in each bin are (*a*) 0.00–0.05: 84; 0.05–0.10: 397; 0.10–0.15: 1,810; 0.15–0.20: 3,146; 0.20–0.25: 2,352; 0.25–0.30: 1,305; 0.30–0.35: 756; 0.35–0.40: 397; >0.40: 360; (*b*) 0.00–0.05: 444; 0.05–0.10: 1,184; 0.10–0.15: 2,473; 0.15–0.20: 2,151; 0.20–0.25: 1,613; 0.25–0.30: 1,117; 0.30–0.35: 740; 0.35–0.40: 444; >0.40: 441.

data). This correlation is strong (rank correlation coefficient of 0.34–0.38) and highly significant even for the original unbinned data (see fig. 3 legend). The GC-RMA expression data set gave similar results (Fig. S4, Supplementary Material online). Because the correlation between r and τ is much higher than that between τ and S, we conclude that the former correlation is not due to the latter. In other words, expression level and tissue specificity independently influence the rate of expression profile evolution.

Our finding of the positive correlation between r and τ appears to be opposite of what Yang, Su, and Li (2005) found. They showed that broadly expressed genes have lower rates of gene expression profile evolution than narrowly expressed genes, which was based on the observation of a positive correlation between expression breadth (B) and the ECI between human and mouse orthologs. Their results may not reflect biological reality for the following three reasons.

First, as aforementioned, they used a potentially problematic approach of applying a signal cutoff to the microarray data and defining expression breadth by counting the number of tissues in which a gene is expressed. Figure 4*a* gives an example illustrating its flaws. It is common that on a microarray chip there are more than one probe set to represent a gene. Theoretically, different probe sets of the same gene should give similar values of τ (or B) because these different probe sets target the same mRNA. However, when the cutoff value of 200 is used for the two probe sets of human *WASPIP* gene, B is substantively different depending on which probe set is used (probe set #1, $B = 2/26 = 0.077$; probe set #2, $B = 17/26 = 0.654$). Figure 4*a* shows that the two probe sets provide relatively consistent expression patterns except that probe set #1 has much lower affinity to the target mRNA than probe set #2. Contrary to B, similar τ values were obtained using these two probe sets (probe set #1, 0.351; probe set #2, 0.334), illustrating that τ is a better measure than B.

Second, because the number of tissues in which a gene is expressed (N) is highly dependent on the signal cutoff used and because ECI is computed from N, one can expect that ECI is also problematic. For example, in figure 4*a*, although the two probe sets represent the same human gene (*WASPIP*) and have congruent expression patterns, the ECI value is low (0.250). In figure 4*b*, although human and mouse *NEU1* genes have substantively different expression profiles, ECI is high (0.961). Contrary to ECI, Pearson's r between expression profiles seems a better index reflecting biological facts (fig. 4*a*, $r = 0.849$; fig. 4*b*, $r = 0.288$).

Finally, because both ECI and B are computed from N, it is expected that ECI and B are not independent from each other. From equation (3), we expect that human-mouse orthologs with larger N should have higher ECI values because by chance they have more opportunities to overlap in expression. To demonstrate this effect, we randomly paired human and mouse genes. As shown in figure 5*a*, the randomly paired genes still show positive correlation between ECI and B, suggesting that the previously observed correlation in Yang, Su, and Li (2005) may not be due to true biological relationships but rather an artifact caused by the dependence between the two parameters used. By contrast, such a correlation does not exist for randomly paired genes when we use τ to measure tissue specificity and r to measure expression profile similarity (fig. 5*b*). Yang, Su, and Li (2005) attempted to avoid the dependence between ECI and B by using different sets of tissues to compute ECI and B. They suggested that their result still holds after this consideration, as shown in their table 1. However, they did not control for the expression level S. Because expression breadth B and mean S are highly correlated (Spearman's rank correlation $= 0.86$, $P < 10^{-300}$ in our data), their observation of conservation of broadly expressed genes could be due to the fact that (1) broadly expressed genes tend to have high expression and (2) highly expressed genes tend to be conserved (fig. 2). The advantage of using τ instead of B is that τ and S are uncorrelated (see above).

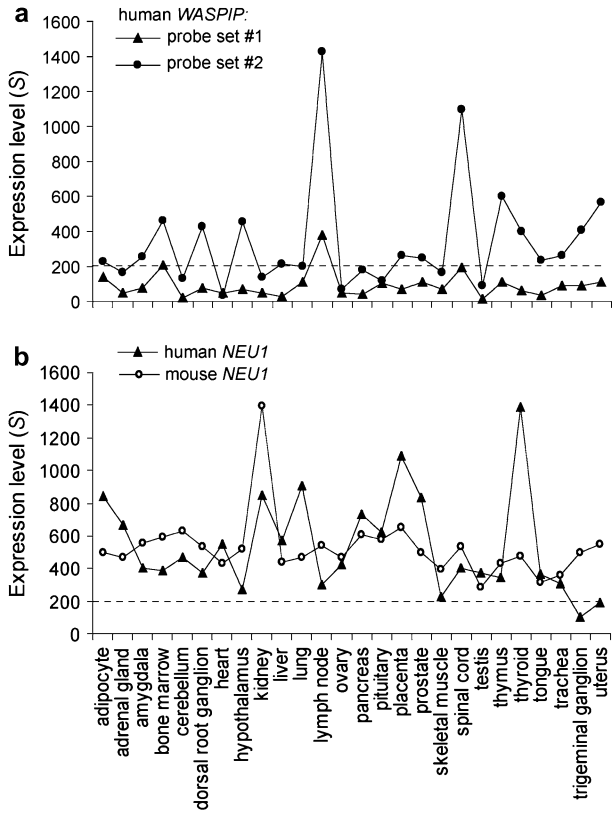

FIG. 4.—Two examples of expression profiles obtained from Gene Atlas V2. (*a*) Profiles of two probe sets (probe set #1, 202663_at; probe set #2, 202664_at) of human *WASPIP* gene. Expression breadth (B) for the probe set #1 and probe set #2 is 0.077 and 0.654, respectively. Tissue specificity (τ) for the two probe sets is 0.351 and 0.334, respectively. For the similarity between the two profiles generated by the two probe sets, ECI = 0.250 and $r = 0.849$. (*b*) Expression profiles of human *NEU1* gene (probe set, 208926_at) and its mouse ortholog (probe set, gnf1m23979_at). The ECI value between the profiles of human-mouse *NEU1* orthologs is 0.961, while r is 0.288.

Previous molecular evolutionary studies have considered the differences between housekeeping and non-housekeeping genes (e.g., Zhang and Li 2004). Housekeeping genes are those expressed in the majority of tissues. It is expected that housekeeping genes have lower tissue specificity than non-housekeeping genes. If one defines human housekeeping genes by $S \geq 200$ in at least 70 of the 73 examined tissues, τ is 0.168 ± 0.001 (mean $\pm$ standard error

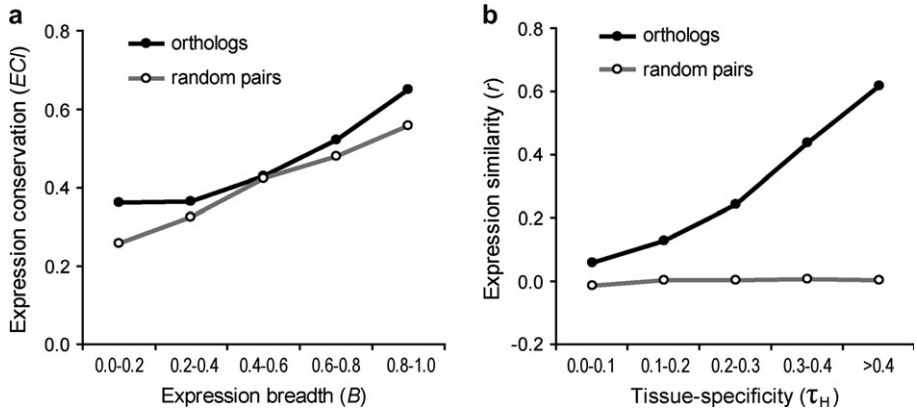

FIG. 5.—The correlation between expression breadth (*B*) and ECI is due to the intrinsic dependence between the two parameters. (*a*) *B* and ECI are positively correlated in both real orthologs and randomly paired human and mouse genes. Following the procedure that Yang, Su, and Li (2005) used to generate their fig. 3, we calculated *B* from the 47 human tissues that are not studied in mouse. (*b*) Tissue specificity (τ) and expression profile similarity (*r*) are positively correlated in real orthologs but not in randomly paired human and mouse genes.

of mean) for the 2,262 housekeeping genes but 0.225 ± 0.001 for the other 8,345 genes, consistent with the above expectation. However, the average expression level is much higher for housekeeping genes (1,351 ± 33) than for the other genes (413 ± 5). Interestingly, we found that the expression profile similarity (*r*) between human-mouse orthologs does not differ between housekeeping genes (0.211 ± 0.006) and the other genes (0.215 ± 0.003). Apparently, high expression levels and low tissue specificities offset each other so that housekeeping genes do not differ from other genes in *r*. We note that although housekeeping genes tend to have low variations in expression level across tissues, the variance is not 0. Furthermore, the relative expression levels across tissues may not be important to housekeeping genes. This may explain why *r* is not higher for housekeeping genes than for non-housekeeping genes.

Similarities and Differences Between Coding Sequence and Expression Profile Evolution

In this work, we used statistical correlations to identify factors that might influence the evolution of expression profiles of mammalian genes. It is important to address (1) whether two quantities are significantly correlated and (2) how strong the correlation is. The important correlations on which our main conclusions are based range from 0.17 to 0.38. These correlations are not particularly high, though statistically highly significant. The relatively low correlations may reflect two facts. First, the evolutionary rate of gene expression profile is determined by multiple factors, each of which may only have a small effect. Second, microarray expression data are known to be noisy, which reduces correlations. Because the evolution of gene expression profiles is poorly understood, it is important to first identify all relevant determinants before one can evaluate their relative contributions. It is also useful to compare the magnitudes of the newly identified correlations with those of well-established correlations, as will be discussed below.

By analyzing over 10,000 human-mouse orthologous gene pairs, we found that highly expressed genes have lower rates of evolution than lowly expressed genes in both coding sequence and expression profile (fig. 6). Gene expression level (*S*) is thought to be the single most important determinant of the rate of coding sequence evolution (Drummond, Raval, and Wilke 2006). We found that the correlation (0.17) between expression profile similarity (*r*) and *S* is slightly higher than that (0.14–0.16) between d_N (or d_N/d_S) and *S* for mammalian genes, suggesting similar importance of expression level in determining the rate of expression profile evolution and the rate of coding sequence evolution.

Do the similar impacts of gene expression level on coding sequence and expression profile evolution suggest a common evolutionary mechanism? A recent study proposed that highly expressed proteins are under stronger pressures to avoid misfolding caused by translational errors; consequently, these proteins have more rigid requirements for their sequences and are more conserved in evolution (i.e., the translational robustness hypothesis) (Drummond et al. 2005). Although this hypothesis may explain why highly expressed genes have low rates of coding sequence evolution, it cannot explain why they also have low rates of expression profile evolution because there is

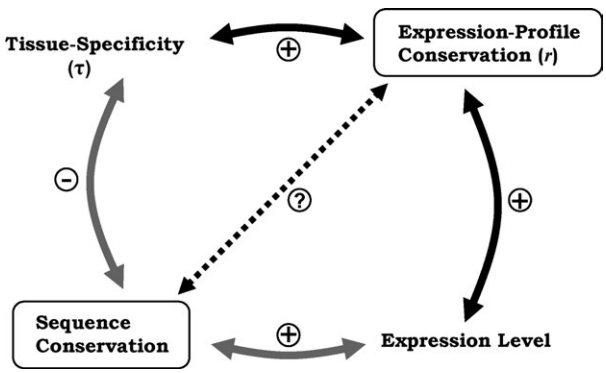

FIG. 6.—A summary of the correlations discussed in this paper. The "+" symbol denotes a positive correlation, while the "−" symbol denotes a negative correlation. The correlations found in previous studies and confirmed in the present work are presented as gray arrows, while those newly found in this study are presented as black arrows. The relationship between the evolutionary conservation of coding sequences and that of expression profiles is unclear.

no link between expression profile conservation and protein misfolding. It has also been proposed that highly expressed genes are functionally more important and therefore are more conserved in their coding sequences (Rocha and Danchin 2004). This functional importance hypothesis may explain our observations if functionally important genes are under strong purifying selection in both coding sequences and expression profiles. However, the functional importance hypothesis was not supported in a previous study of yeasts (Drummond et al. 2005). Furthermore, in yeasts and bacteria, only a small fraction of the strong correlation between gene expression level and d_N may be explained by their covariations with gene importance, which is measured by the fitness reduction caused by gene deletions (Rocha and Danchin 2004; Zhang and He 2005). The main reason behind the impact of gene expression level on the rate of coding sequence evolution is still unclear. It is possible that the apparently similar influences of gene expression level on coding sequence divergence and expression profile divergence have different underlying causes.

We found that tissue specificity has opposite impacts on the rate of coding sequence evolution and the rate of expression profile evolution. Compared with a gene with low tissue specificity, a gene with high tissue specificity tends to evolve faster in its coding sequence but slower in its expression profile (fig. 6). It has been suggested that there is less functional constraint on a protein sequence if it is expressed only in a small number of tissues (Duret and Mouchiroud 2000). At the same time, tissue-specific genes may be more adaptable due to fewer pleiotropic effects (Duret and Mouchiroud 2000). As a consequence, tissue specificity and d_N become positively correlated. More detailed causal effects regarding this relationship have been discussed in Zhang and Li (2004). However, it is worth noting that the correlation between tissue specificity (τ) and d_N is low (Spearman's rank correlation = 0.089) in our analysis. Previous studies demonstrating an impact of tissue specificity on coding sequence evolution were likely confounded by the influence of expression level as expression cutoffs were used to define tissue specificity (Duret and Mouchiroud 2000; Zhang and Li 2004). In the present study, however, the impact of tissue specificity can be clearly separated as τ is uncorrelated with expression level.

The correlation between expression profile similarity and τ ranges from 0.34 to 0.38, indicating that the impact of tissue specificity on expression profile evolution is much greater than that on coding sequence evolution. Given the large estimation error of expression profile similarity caused by microarray technologies (Liao and Zhang 2006), the high correlation observed prompts us to believe that tissue specificity is one of the most important determinants of the evolutionary rate of gene expression profile in mammals. Why do highly tissue-specific genes have a low rate of expression profile evolution? It is possible that for a tissue-specific gene, its function is highly specialized for the tissues where it is expressed. Expression of the gene in a different tissue would make the protein physiologically useless or even detrimental. In contrast, non–tissue-specific genes may be more tolerant to changes of expression level in various tissues, thus having relatively high rates of expression profile evolution. Taken together, expression profile evolution and coding sequence evolution appear to be governed by different principles.

A recent study based on human-chimpanzee comparisons suggested that the evolutionary rate of the expression level of a gene is positively correlated with the evolutionary rate of its coding sequence (Khaitovich et al. 2005). However, it is unclear whether the evolutionary rate of expression profile is correlated with that of coding sequence (fig. 6). Several studies using human-mouse orthologs do not find such a correlation (Jordan et al. 2004; Yanai, Graur, and Ophir 2004; Jordan, Marino-Ramirez, and Koonin 2005). Our previous study revealed a weak positive correlation between these two quantities when the Euclidean distance was used to measure the profile similarity of human-mouse orthologs (Liao and Zhang 2006). However, such a correlation was not observed when Pearson's r was used to measure the profile similarity. Figure 6 illustrates that these ambiguous results might be related to the different effects of the expression level and tissue specificity on the evolutionary rate of coding sequence and that of expression profile.

It should be emphasized that genome-wide analysis of gene expression evolution has just begun, and most studies have focused on identifying correlations. When a higher quantity and quality of data become available, the underlying causes of the identified correlations and the relative contributions of various factors may be examined. We also want to stress that the impacts of expression level and tissue specificity on the evolutionary rate of expression profile that we report in this work should be confirmed in other data sets and other species. Unlike the study of gene/protein sequence evolution, in which various evolutionary distances have been developed (Li 1997; Nei and Kumar 2000), the study of expression profile divergence still lacks a good distance measure. All the distances so far introduced (r, Euclidean distance, and ECI) only measure the relative divergence of expression profiles but tell nothing about the number of genetic changes that are responsible for the expression divergence. Understanding the molecular genetic mechanisms of expression regulation will facilitate the development of such distance measures, which will in turn help elucidate the mode and cause of expression evolution.

Supplementary Material

Supplementary figures S1–S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Luis Chaves, Wendy Grus, Xionglei He, Ondrej Podlaha, and two anonymous reviewers for valuable comments. This work was supported by research grants from University of Michigan and National Institutes of Health to J.Z.

Literature Cited

Carroll, S. B. 2005. Evolution at two levels: on genes and form. PLoS Biol. **3**:e245.

Cavalieri, D., J. P. Townsend, and D. L. Hartl. 2000. Manifold anomalies in gene expression in a vineyard isolate of

Saccharomyces cerevisiae revealed by DNA microarray analysis. Proc. Natl. Acad. Sci. USA **97**:12369–12374.
Denver, D. R., K. Morris, J. T. Streelman, S. K. Kim, M. Lynch, and W. K. Thomas. 2005. The transcriptional consequences of mutation and natural selection in Caenorhabditis elegans. Nat. Genet. **37**:544–548.
Drummond, D. A., J. D. Bloom, C. Adami, C. O. Wilke, and F. H. Arnold. 2005. Why highly expressed proteins evolve slowly. Proc. Natl. Acad. Sci. USA **102**:14338–14343.
Drummond, D. A., A. Raval, and C. O. Wilke. 2006. A single determinant dominates the rate of yeast protein evolution. Mol. Biol. Evol. **23**:327–337.
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**:68–74.
Enard, W., P. Khaitovich, J. Klose et al. (13 co-authors). 2002. Intra- and interspecific variation in primate gene expression patterns. Science **296**:340–343.
Gu, X., Z. Zhang, and W. Huang. 2005. Rapid evolution of expression and regulatory divergences after yeast gene duplication. Proc. Natl. Acad. Sci. USA **102**:707–712.
Gu, Z., D. Nicolae, H. H. Lu, and W. H. Li. 2002. Rapid divergence in expression between duplicate genes inferred from microarray data. Trends Genet. **18**:609–613.
Gu, Z., S. A. Rifkin, K. P. White, and W. H. Li. 2004. Duplicate genes increase gene expression diversity within and between species. Nat. Genet. **36**:577–579.
Hastings, K. E. 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.
He, X., and J. Zhang. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics **169**:1157–1164.
Hill, A. A., E. L. Brown, M. Z. Whitley, G. Tucker-Kellogg, C. P. Hunter, and D. K. Slonim. 2001. Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls. Genome Biol. **2**:research0055.
Hubbell, E., W. M. Liu, and R. Mei. 2002. Robust estimators for expression analysis. Bioinformatics **18**:1585–1592.
Huminiecki, L., and K. H. Wolfe. 2004. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. Genome Res. **14**:1870–1879.
Irizarry, R. A., B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics **4**:249–264.
Jordan, I. K., L. Marino-Ramirez, and E. V. Koonin. 2005. Evolutionary significance of gene expression divergence. Gene **345**:119–126.
Jordan, I. K., L. Marino-Ramirez, Y. I. Wolf, and E. V. Koonin. 2004. Conservation and coevolution in the scale-free human gene coexpression network. Mol. Biol. Evol. **21**:2058–2070.
Kasprzyk, A., D. Keefe, D. Smedley, D. London, W. Spooner, C. Melsopp, M. Hammond, P. Rocca-Serra, T. Cox, and E. Birney. 2004. EnsMart: a generic system for fast and flexible access to biological data. Genome Res. **14**:160–169.
Khaitovich, P., I. Hellmann, W. Enard, K. Nowick, M. Leinweber, H. Franz, G. Weiss, M. Lachmann, and S. Paabo. 2005. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. Science **309**:1850–1854.
Khaitovich, P., G. Weiss, M. Lachmann, I. Hellmann, W. Enard, B. Muetzel, U. Wirkner, W. Ansorge, and S. Paabo. 2004. A neutral model of transcriptome evolution. PLoS Biol. **2**: 682–689.
King, M. C., and A. C. Wilson. 1975. Evolution at two levels in humans and chimpanzees. Science **188**:107–116.
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.
Li, W.-H. 1997. Molecular evolution. Sinauer Associates, Sunderland, Mass.
Liao, B.-Y., and J. Zhang. 2006. Evolutionary conservation of expression profiles between human and mouse orthologous genes. Mol. Biol. Evol. **23**:530–540.
Makova, K. D., and W. H. Li. 2003. Divergence in the spatial pattern of gene expression between human duplicate genes. Genome Res. **13**:1638–1645.
Murphy, W. J., P. A. Pevzner, and S. J. O'Brien. 2004. Mammalian phylogenomics comes of age. Trends Genet. **20**:631–639.
Nei, M., and S. Kumar. 2000. Molecular evolution and phylogenetics. Oxford University Press, New York.
Oleksiak, M. F., G. A. Churchill, and D. L. Crawford. 2002. Variation in gene expression within and among natural populations. Nat. Genet. **32**:261–266.
Pal, C., B. Papp, and L. D. Hurst. 2001. Highly expressed genes in yeast evolve slowly. Genetics **158**:927–931.
Ponger, L., L. Duret, and D. Mouchiroud. 2001. Determinants of CpG islands: expression in early embryo and isochore structure. Genome Res. **11**:1854–1860.
Ranz, J. M., C. I. Castillo-Davis, C. D. Meiklejohn, and D. L. Hartl. 2003. Sex-dependent gene expression and evolution of the Drosophila transcriptome. Science **300**:1742–1745.
Rifkin, S. A., D. Houle, J. Kim, and K. P. White. 2005. A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. Nature **438**:220–223.
Rifkin, S. A., J. Kim, and K. P. White. 2003. Evolution of gene expression in the Drosophila melanogaster subgroup. Nat. Genet. **33**:138–144.
Rocha, E. P., and A. Danchin. 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol. Biol. Evol. **21**:108–116.
Schug, J., W. P. Schuller, C. Kappen, J. M. Salbaum, M. Bucan, and C. J. Stoeckert, Jr. 2005. Promoter features related to tissue specificity as measured by Shannon entropy. Genome Biol. **6**:R33.
Springer, M. S., W. J. Murphy, E. Eizirik, and S. J. O'Brien. 2003. Placental mammal diversification and the Cretaceous-Tertiary boundary. Proc. Natl. Acad. Sci. USA **100**:1056–1061.
Su, A. I., M. P. Cooke, K. A. Ching et al. (14 co-authors). 2002. Large-scale analysis of the human and mouse transcriptomes. Proc. Natl. Acad. Sci. USA **99**:4465–4470.
Su, A. I., T. Wiltshire, S. Batalov et al. (13 co-authors). 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc. Natl. Acad. Sci. USA **101**:6062–6067.
Subramanian, S., and S. Kumar. 2004. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics **168**:373–381.
Vinogradov, A. E. 2004. Compactness of human housekeeping genes: selection for economy or genomic design? Trends Genet. **20**:248–253.
Wall, D. P., A. E. Hirsh, H. B. Fraser, J. Kumm, G. Giaever, M. B. Eisen, and M. W. Feldman. 2005. Functional genomic analysis of the rates of protein evolution. Proc. Natl. Acad. Sci. USA **102**:5483–5488.
Winter, E. E., L. Goodstadt, and C. P. Ponting. 2004. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. Genome Res. **14**:54–61.
Wu, Z., R. A. Irizarry, R. Gentleman, F. Martinez Murillo, and F. Spencer. 2004. A model based background adjustment for oligonucleotide expression arrays. J. Am. Stat. Assoc. **99**:909–917.
Yanai, I., H. Benjamin, M. Shmoish et al. (12 co-authors). 2005. Genome-wide midrange transcription profiles reveal

expression level relationships in human tissue specification. Bioinformatics **21**:650–659.

Yanai, I., D. Graur, and R. Ophir. 2004. Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. OMICS **8**:15–24.

Yang, J., A. I. Su, and W.-H. Li. 2005. Gene expression evolves faster in narrowly than in broadly expressed mammalian genes. Mol. Biol. Evol. **22**:2113–2118.

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

Zhang, J., and X. He. 2005. Significant impact of protein dispensability on the instantaneous rate of protein evolution. Mol. Biol. Evol. **22**:1147–1155.

Zhang, L., and W. H. Li. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Mol. Biol. Evol. **21**:236–239.

Takashi Gojobori, Associate Editor

Accepted March 1, 2006