

# Impacts of Gene Essentiality, Expression Pattern, and Gene Compactness on the Evolutionary Rate of Mammalian Proteins

Ben-Yang Liao,\* Nicole M. Scott,† and Jianzhi Zhang\*

\*Department of Ecology and Evolutionary Biology, University of Michigan and †Department of Human Genetics, University of Michigan

Understanding the determinants of the rate of protein sequence evolution is of fundamental importance in evolutionary biology. Many recent studies have focused on the yeast because of the availability of many genome-wide expressional and functional data. Yeast studies revealed a predominant role of gene expression level and a minor role of gene essentiality in determining the rate of protein sequence evolution. Whether these rules apply to complex organisms such as mammals is unclear. Here we assemble a list of 1,138 essential and 2,341 nonessential mouse genes based on targeted gene deletion experiments and report a significant impact of gene essentiality on the rate of mammalian protein evolution. Gene expression level has virtually no effect, although tissue specificity in expression pattern has a strong influence. Unexpectedly, gene compactness, measured by average intron size and untranslated region length, has the greatest influence. Hence, the relative importance of the various factors in determining the rate of mammalian protein evolution is gene compactness > gene essentiality  $\approx$  tissue specificity > expression level. Our results suggest a considerable variation in rate determinants between unicellular organisms such as the yeast and multicellular organisms such as mammals.

## Introduction

What determines the rate of protein sequence evolution is a fundamental question in molecular evolution. It is well known that the evolutionary rates of different proteins in a genome vary by several orders of magnitude (Dayhoff 1972; Li 1997). This variation is typically explained by differences in the mutation rate and selection intensity among genes (Kimura 1983; Li 1997). However, the biological factors underlying such differences had not been examined with sufficiently large data until a few years ago, when genome sequences and functional genomic data became available. Factors that have been shown to influence the protein evolutionary rate include gene essentiality (Hirsh and Fraser 2001; Jordan et al. 2002; Wall et al. 2005; Zhang and He 2005), gene expression level (Pal et al. 2001b; Akashi 2003; Rocha and Danchin 2004; Subramanian and Kumar 2004; Drummond et al. 2006), tissue specificity (Hastings 1996; Duret and Mouchiroud 2000; Subramanian and Kumar 2004; Winter et al. 2004; Zhang and Li 2004), presence of a duplicate copy (Nembaware et al. 2002; Castillo-Davis and Hartl 2003; Yang et al. 2003), properties in the protein interaction network (Fraser et al. 2002; Fraser 2005; Hahn and Kern 2005; Makino and Gojobori 2006), local recombination rate (Pal et al. 2001b), and pleiotropy (He and Zhang forthcoming), although some of these factors are interrelated. In the past few years, many studies have focused on unicellular organisms, particularly the yeast *Saccharomyces cerevisiae*, due to the early availability of many functional genomic data for this model organism. With the advancement of mammalian genomics, it becomes possible to conduct genome-wide analysis of several biological factors that potentially influence the rate of mammalian protein evolution and to compare the relative importance of these factors in yeasts and mammals, respective representatives of unicellular and multicellular eukaryotes.

Key words: evolutionary rate, gene essentiality, tissue specificity, expression level, gene compactness, mammals.

E-mail: jianzhi@umich.edu.

*Mol. Biol. Evol.* 23(11):2072–2080. 2006

doi:10.1093/molbev/msl076

Advance Access publication August 3, 2006

Among the potential rate determinants, gene essentiality is perhaps the most studied and debated factor. Essential genes refer to those that cause lethality or infertility when deleted. Based on the neutral theory of molecular evolution (Kimura and Ohta 1974), it was predicted that essential genes are subject to stronger selective constraints and, therefore, evolve more slowly than nonessential genes (Wilson et al. 1977). However, Hurst and Smith (1999) failed to verify this prediction when they compared 67 essential genes and 108 nonessential genes of the mouse. Although subsequent analysis of bacterial and yeast genes found gene essentiality to be an important rate determinant (Hirsh and Fraser 2001; Jordan et al. 2002), these results were suggested to arise from a confounding factor of the gene expression level (Pal et al. 2003; Rocha and Danchin 2004). More recent analyses, however, showed that gene essentiality has a small, yet statistically significant, impact on the evolutionary rate of yeast proteins even when the gene expression level is controlled for (Wall et al. 2005; Zhang and He 2005). Nonetheless, despite the availability of many mouse strains produced in targeted gene deletion experiments, whether gene essentiality influences mammalian protein evolution remains unsolved due to the lack of a comprehensive list of essential and nonessential genes.

The importance of gene expression level in determining the protein evolutionary rate in yeasts and bacteria is well established (Pal et al. 2001a; Rocha and Danchin 2004; Zhang and He 2005; Drummond et al. 2006), although the molecular evolutionary mechanisms are unclear and debated (Akashi 2003; Drummond et al. 2005). Unlike unicellular organisms, mammalian cells are highly differentiated, and different types of cells turn on different sets of genes to maintain their identities and functions. Hence, both the expression level and tissue specificity of expression may be important in determining the rate of mammalian gene evolution. In fact, previous studies of mammalian genes showed higher evolutionary rates among lowly expressed genes than highly expressed genes (Subramanian and Kumar 2004) and higher rates among tissue-specific genes than housekeeping genes (Duret and Mouchiroud 2000; Winter et al. 2004; Zhang and Li 2004). However, because housekeeping genes tend to be highly expressed (Vinogradov 2004; Liao and Zhang 2006a), it is unknown whether

expression level and tissue specificity have independent influences on the evolutionary rate.

A previous study of 363 mouse and rat genes showed a significant, but weak, negative correlation between protein length and the rate of protein sequence evolution (Zhang 2001). An opposite pattern, however, was found in the fruit fly (Lemos et al. 2005). Recent studies also showed that highly expressed genes tend to code for short proteins and have short introns (Castillo-Davis et al. 2002). Because highly expressed genes tend to have low rates of protein evolution, one would expect a positive correlation between protein (or intron) length and the rate of protein evolution. It is interesting to test this prediction.

In the present study, we first compile a list of 3,479 mouse genes with essentiality information derived from targeted gene deletion data. We then study the influences of gene essentiality, gene expression level, tissue specificity, and gene compactness (in terms of protein length, average intron length, and untranslated region (UTR) length) on the rate of mammalian protein evolution. We conduct a series of partial correlation analyses to disentangle the contributions of various factors and compare our results with findings from the yeast. Our results reveal a great variation in rate determinants between unicellular and multicellular organisms.

## Materials and Methods

### Mouse Essential and Nonessential Genes

Mouse genes subject to targeted deletion experiments were downloaded from Mouse Genome Database (MGD) (<http://www.informatics.jax.org/>). Only those genes having one corresponding Ensembl gene name were kept for subsequent analysis. These genes were classified into essential and nonessential genes based on their knockout phenotypic codes (MP numbers) provided by MGD. By definition, essential genes are those with the knockout phenotype of lethality or sterility. That is, those entries possessing embryonic lethality (MP: 0002080), prenatal lethality (MP: 0002081), survival postnatal lethality (MP: 0002082), premature death or induced morbidity (MP: 0002083), or reproductive system phenotype (MP: 0002161) were grouped as essential genes. All other genes associated with a phenotypic classification term, including those entries with a normal phenotype, were grouped as nonessential genes. The primary data set included 1,138 essential and 2,341 nonessential mouse genes.

### Gene Orthology and Evolutionary Rate

The homology information of mouse and rat genes was obtained from Ensembl EnsMart (<http://www.ensembl.org/Multi/martview>). There were several annotated homology relationships between mouse and rat 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. The number of synonymous substitutions per synonymous site ( $d_S$ ) and the number of nonsynonymous substitutions per nonsynonymous site ( $d_N$ ) between mouse and rat orthologs were estimated

by the maximum likelihood method of Yang (1997) and retrieved from Ensembl EnsMart.

### Structural and Functional Annotations of Mouse Genes

The structural and functional annotations of mouse genes were obtained from Ensembl version 31. Chromosomal positions, coding sequence (CDS) lengths, intron numbers, intron lengths, and 5' and 3' UTR lengths of mouse genes were retrieved from Ensembl EnsMart (<http://www.ensembl.org/Multi/martview>) (Kasprzyk et al. 2004). For alternatively spliced genes, we chose structural information of the splice form with the longest CDS. Genes having immune-related functions were identified from the Gene Ontology description (<http://www.geneontology.org/>) contained in Ensembl database. It should be noted that not all mouse genes in the preliminary data set have rat orthologs. After removing mouse genes without UBRH rat orthologs, 1,038 essential and 2,126 nonessential mouse genes were kept for subsequent analysis.

The gene structure annotation of the yeast *S. cerevisiae* was also obtained from Ensembl EnsMart. Nucleotide substitution rates between *S. cerevisiae* and *Saccharomyces bayanus* orthologous genes were obtained from Zhang and He (2005).

### Analysis of Gene Expression Pattern

The spatial expression information of mouse genes was obtained from the Gene Atlas V2 data set (<http://symatlas.gnf.org/SymAtlas/>). This data set was generated by hybridization of RNAs from 61 mouse tissues onto Affymetrix microarray chips (GNF1M) (Su et al. 2004). To assign expression data from probe sets to corresponding Ensembl mouse genes, we aligned probe sequences of each probe set to the Ensembl cDNA sequences (Mus\_musculus.NCBIM33.feb.cdna.fa; <http://www.ensembl.org/info/data/download.html>) using BlastN (<http://www.ncbi.nlm.nih.gov/blast/>). Only those probe sets in which all 10 matching probes perfectly matched with the same Ensembl gene were considered to be valid. The expression level detected by each probe set was obtained as the signal intensity ( $S$ ) computed from MAS 5.0 algorithm (MAS5) (Hubbell et al. 2002). The  $S$  values were averaged among replicates.

In the present study, we measured 2 properties of the mouse gene expression pattern: expression level ( $ExpLev$ ) and tissue specificity ( $\tau$ ).  $ExpLev$  is defined as the average signal intensity ( $S$ ) of a mouse gene across 61 examined tissues. The tissue specificity of a gene is defined as the heterogeneity of its expression level across all the tissues and is

estimated by  $\tau = \frac{\sum_{j=1}^n \left(1 - \frac{\log_2 S(j)}{\log_2 S_{\max}}\right)}{n-1}$ , where  $n = 61$  is the number of mouse tissues examined here and  $S_{\max}$  is the highest expression signal of the gene across all tissues (Yanai et al. 2005). To minimize the influence of noise from low intensities, we arbitrarily let  $S(j)$  be 100 if it is lower than 100 (Liao and Zhang 2006a). The  $\tau$  value ranges from 0 to 1, with higher values indicating greater variations in expression level across tissues and thus higher tissue specificity. The advantage of using  $\tau$  rather than expression breadth, which requires an arbitrary cutoff to determine

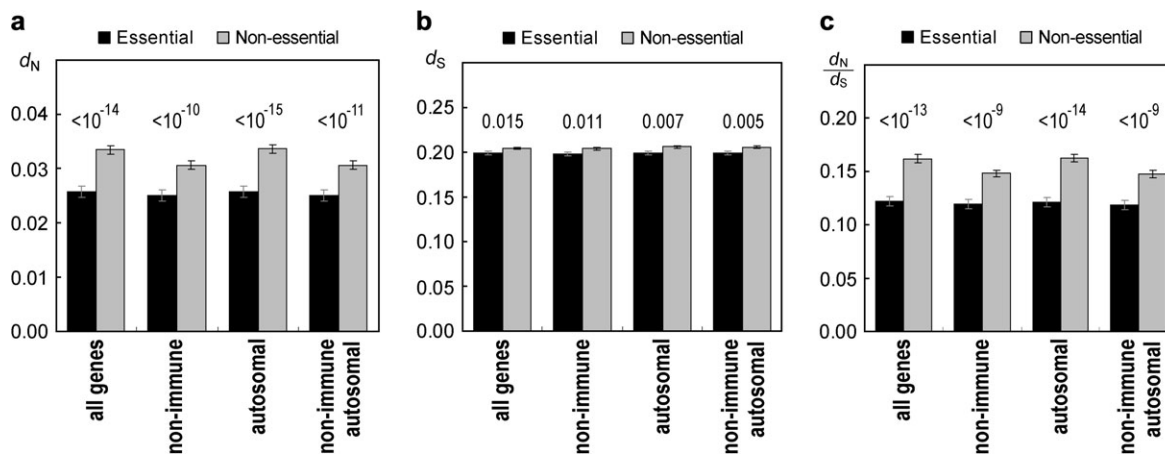


FIG. 1.—Nonessential mouse genes evolve faster than essential genes. Average mouse–rat (a)  $d_N$ , (b)  $d_S$ , and (c)  $d_N/d_S$  values of essential and nonessential genes are shown.  $P$  value from the test of the null hypothesis of no difference between essential and nonessential genes is shown above each comparison (Mann–Whitney  $U$  test). Error bars represent the standard error of the mean. All genes: 1,038 essential and 2,126 nonessential. Non-immune system genes: 1,006 essential and 1,977 nonessential. Autosomal genes: 1,014 essential and 2,053 nonessential. Autosomal, non-immune system genes: 982 essential and 1,908 nonessential.

whether a gene is expressed in a given tissue, has been extensively discussed (Liao and Zhang 2006a). Some mouse genes are represented by more than one probe set on the microarray. Because it was not possible to tell which probe set provides the best expression measure of a target gene (Liao and Zhang 2006b), we computed *ExpLev* and  $\tau$  by averaging the values derived from the different probe sets of the same gene. The final data set used in partial correlation analyses contained 2,575 mouse genes with knockout phenotypes, orthologous rat genes, and structural and expression data. Among them, 852 were essential and 1,723 were nonessential.

## Results

### Nonessential Proteins Evolve Faster than Essential Proteins

We compiled a list of 1,138 essential and 2,341 non-essential genes using mouse targeted gene deletion data. Among them, 1,038 essential and 2,126 nonessential genes have orthologous genes in the rat. The number of synonymous substitutions per synonymous site ( $d_S$ ) and the number of nonsynonymous substitutions per nonsynonymous site ( $d_N$ ) were estimated for these genes using mouse and rat orthologs. We found a significant difference between essential and nonessential genes in  $d_N$  ( $P < 10^{-14}$ , Mann–Whitney  $U$  test; fig. 1a). On average,  $d_N$  is 30% greater for nonessential genes than essential genes. We noticed that X-linked genes and immune system genes are slightly overrepresented in the nonessential group (3.3% and 7.0%), compared with the essential group (2.3% and 3.1%). Because X-linked mammalian genes may behave differently from autosomal genes due to differences in gene content, mutation rate, and selection intensity (Wang et al. 2001; Malcom et al. 2003; Lu and Wu 2005) and immune system genes tend to be under diversifying positive selection (Hughes and Nei 1988; Hughes 1999), we repeated the above analysis by removing X-linked genes and immune-related genes. Our results, however, remain virtually unchanged (fig. 1a). Although  $d_S$  is also significantly higher

for nonessential genes than essential genes, the difference in mean  $d_S$  between the 2 groups is small ( $\sim 3\%$ ) (fig. 1b). The average  $d_N/d_S$  ratio of nonessential genes is 24–34% greater than that of essential genes, depending on whether X-linked genes and immune system genes are considered or not (fig. 1c). Thus, there is a significantly negative correlation between gene essentiality and  $d_N/d_S$  (or  $d_N$ ) (table 1). These results indicate that gene essentiality affects the rate of mammalian protein evolution by influencing the selective constraint on the proteins.

### Effects of Gene Expression Level and Tissue Specificity on the Rate of Protein Evolution

Two gene expression properties, expression level (Pal et al. 2001a; Rocha and Danchin 2004; Subramanian and Kumar 2004; Zhang and He 2005; Drummond et al. 2006) and tissue specificity (Hastings 1996; Duret and Mouchiroud 2000; Subramanian and Kumar 2004; Zhang and Li 2004), have been shown to affect the rate of protein sequence evolution to various degrees in different species. Specifically, highly expressed genes and non-tissue-specific genes tend to evolve slowly. Analysis based on our data set confirms these findings (table 1 and fig. 2). Interestingly, although gene expression level is the most important rate determinant in bacteria (Rocha and Danchin 2004) and yeast (Drummond et al. 2006), the correlation between gene expression level (*ExpLev*) and  $d_N$  is weak (Spearman's  $\rho = -0.04$ ) and only marginally significant ( $P = 0.041$ ) in mammals. Similar results are obtained when essential and nonessential genes are analyzed separately. By contrast, the correlation between tissue specificity ( $\tau$ ) and  $d_N$  is much stronger ( $\rho = 0.168$ ,  $P < 10^{-16}$ ). We noticed that tissue-specific genes not only have greater  $d_N/d_S$  but also greater  $d_S$  values (fig. 2), implying that faster protein evolution of tissue-specific genes may have resulted from both higher mutation rate and lower purifying selection. Because average  $d_S$  does not exhibit the same magnitude of increase as average  $d_N$  while  $\tau$  becomes larger ( $\sim 17\%$  increase vs.  $\sim 90\%$  increase), mutation rate

**Table 1**  
**Spearman's Rank Correlation Coefficient ( $\rho$ ) between Various Factors and  $d_N$ ,  $d_S$ , or  $d_N/d_S$** 

	$d_N$		$d_S$		$d_N/d_S$	
	$\rho$	<i>P</i> Value	$\rho$	<i>P</i> Value	$\rho$	<i>P</i> Value
Essentiality	-0.139	$1.36 \times 10^{-12}$	-0.048	$1.56 \times 10^{-2}$	-0.133	$1.45 \times 10^{-11}$
Expression pattern						
<i>ExpLev</i>	-0.040	$4.08 \times 10^{-2}$	0.056	$4.47 \times 10^{-3}$	-0.051	$9.28 \times 10^{-3}$
Tissue specificity ( $\tau$ )	0.168	$1.1 \times 10^{-17}$	0.094	$1.97 \times 10^{-6}$	0.152	$9.42 \times 10^{-15}$
Gene structures						
CDS length	-0.009	$6.31 \times 10^{-1}$	0.066	$8.69 \times 10^{-4}$	-0.032	$1.00 \times 10^{-1}$
UTR length	-0.213	$8.29 \times 10^{-28}$	-0.042	$3.49 \times 10^{-2}$	-0.213	$8.14 \times 10^{-28}$
5'-UTR length	-0.141	$5.67 \times 10^{-13}$	-0.050	$1.17 \times 10^{-2}$	-0.132	$1.78 \times 10^{-11}$
3'-UTR length	-0.106	$6.13 \times 10^{-8}$	-0.010	$6.11 \times 10^{-1}$	-0.111	$1.59 \times 10^{-8}$
Intron number	0.016	$4.29 \times 10^{-1}$	0.087	$1.05 \times 10^{-5}$	-0.013	$5.13 \times 10^{-1}$
Intron length (average)	-0.163	$7.31 \times 10^{-17}$	-0.012	$5.47 \times 10^{-1}$	-0.175	$3.96 \times 10^{-19}$

NOTE.—Essentiality is 1 for essential genes and 0 for nonessential genes. *P* values show the probabilities of the observations under the hypothesis of no correlation. The analysis is based on 2,575 mouse genes and their rat orthologs.

bias is unlikely to be the main cause for high  $d_N$  of tissue-specific genes. Our result is consistent with that of Zhang and Li (2004).

Because the expression level and tissue specificity may be correlated, we measured the partial correlation between *ExpLev* and  $d_N$  by controlling for  $\tau$ . Although the partial correlation becomes stronger and more significant ( $\rho = -0.061$ ,  $P = 0.002$ ), it is still not comparable to the partial correlation between  $\tau$  and  $d_N$  when *ExpLev* is controlled for ( $\rho = 0.174$ ,  $P < 10^{-18}$ ). These results suggest that tissue specificity is much more important than average expression level in determining the rate of mammalian protein sequence evolution.

#### Compact Genes Have High Rates of Evolution

Although a significant positive correlation between the CDS length and  $d_N$  was observed in fruit fly (Lemos et al. 2005) and a significant negative correlation was observed in a set of 363 mouse and rat genes (Zhang 2001), no significant correlation is found in our data (table 1). Surprisingly, we found a negative correlation between UTR length and  $d_N$  (or  $d_N/d_S$ ) (table 1 and fig. 3). For example, the mean  $d_N$  of genes with a total UTR length of  $<300$  nt is about twice that of genes with a total UTR length of  $>2,400$  nt (fig. 3a). Similarly, we found a negative correlation between average size per intron in a gene (but not intron number) and  $d_N$  (or  $d_N/d_S$ ) of the gene (table 1 and fig. 4). The mean  $d_N$  of genes with an average intron size of  $<1,000$  nt is over 5 times that of genes with an average intron size of  $>8,000$  nt (fig. 4a). The correlations between gene compactness and  $d_N$  are of comparable or even higher magnitudes than that between tissue specificity ( $\tau$ ) and  $d_N$  (table 1).

In the above analysis, we used the longest splice form for those genes that have alternative splicing. We repeated the above analysis by using the shortest splice form or by removing genes with alternative splicing. The results are essentially the same (supplementary tables 1 and 2, Supplementary Material online). There are also many overlapping (including nested) genes in the mouse genome (Veeramachaneni et al. 2004). Removing these genes does not change our result (supplementary table 3, Supplementary Material online).

#### Relative Impacts of Gene Essentiality, Expression Pattern, and Gene Compactness on the Evolutionary Rate

The above-examined factors are not completely independent in determining the rate of protein sequence evolution. For instance, genes with high expression levels tend to have small introns ( $\rho = -0.079$ ,  $P < 10^{-4}$ ). In order to separate the contributions of multiple factors, we applied partial correlation analyses. Although a recent study suggested that principle component regression analysis is superior to partial correlation analysis for noisy data (Drummond et al. 2006), subsequent analytical and empirical analyses do not support this view (S.-H. Kim and S. Yi, personal communication). In our partial correlation analysis, we focused on the correlation between the evolution rate and 1 of the 4 factors (i.e., gene essentiality, expression pattern, and gene compactness), by controlling the other 2 factors. All factors having significant effects on the evolutionary rate in table 1 show significant and independent effects on  $d_N$  and  $d_N/d_S$ , with the exception of *ExpLev* (table 2). After controlling for gene essentiality, the negative correlation between *ExpLev* and  $d_N$  is no longer significant ( $P = 0.06$ ) and that between *ExpLev* and  $d_N/d_S$  is only marginally significant ( $P = 0.014$ ), suggesting that the weak negative correlation between gene expression level and protein evolutionary rate in table 1 may be due to the fact that essential genes tend to have both high *ExpLev* and low  $d_N$ . Our result thus suggests that the effect of expression level itself on the evolutionary rate of mammalian proteins is negligible. We notice that genes with high expression levels tend to have high  $d_S$  but low  $d_N/d_S$  (table 2). Hence, the lack of a correlation between *ExpLev* and  $d_N$  may be due to the opposite effects of high mutation rates and strong purifying selection at highly expressed genes. On comparing the effects of gene essentiality, expression level, tissue specificity, and gene compactness, gene compactness seems to have the strongest and most significant impact on the rate of mammalian protein evolution (tables 1 and 2). Based on the partial correlation analysis (table 2), we conclude that the relative importance of the factors in determining the rate of mammalian protein evolution is gene compactness  $>$  gene essentiality  $\approx$  tissue specificity  $>$  gene expression level.

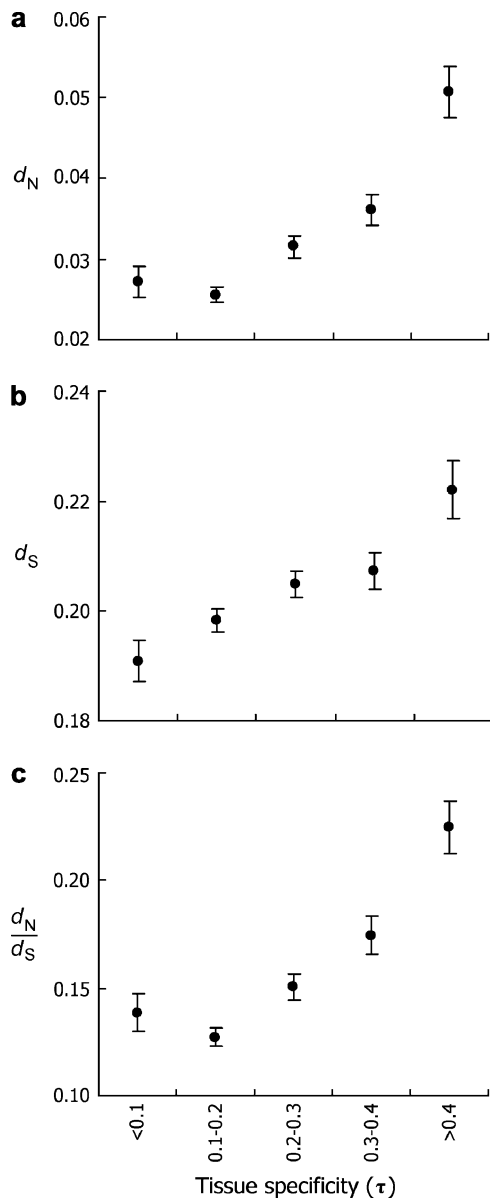


FIG. 2.—Evolutionary rate of mouse genes positively correlates with tissue specificity,  $\tau$ . Average mouse–rat (a)  $d_N$ , (b)  $d_S$ , and (c)  $d_N/d_S$  values of each bin are shown. Error bars represent the standard error of the mean. The numbers of genes in each bin are <0.1: 285; 0.1–0.2: 954; 0.2–0.3: 756; 0.3–0.4: 398; >0.4: 182, with the total number of genes being 2,575.

## Discussion

In this work, we used statistical analysis to study the determinants of the rate of mammalian protein sequence evolution. Because there are potentially many rate determinants and because some measures of these determinants (e.g., gene expression level and tissue specificity) have large estimation errors (Wall et al. 2005; Liao and Zhang 2006b), it is not unexpected that the observed correlation coefficients are not very high. We thus evaluate the impact of each factor by considering both the statistical significance in correlation analysis and the magnitude of the correlation. We also compare the impacts of different factors for a given species and the impacts of the same factor across species.

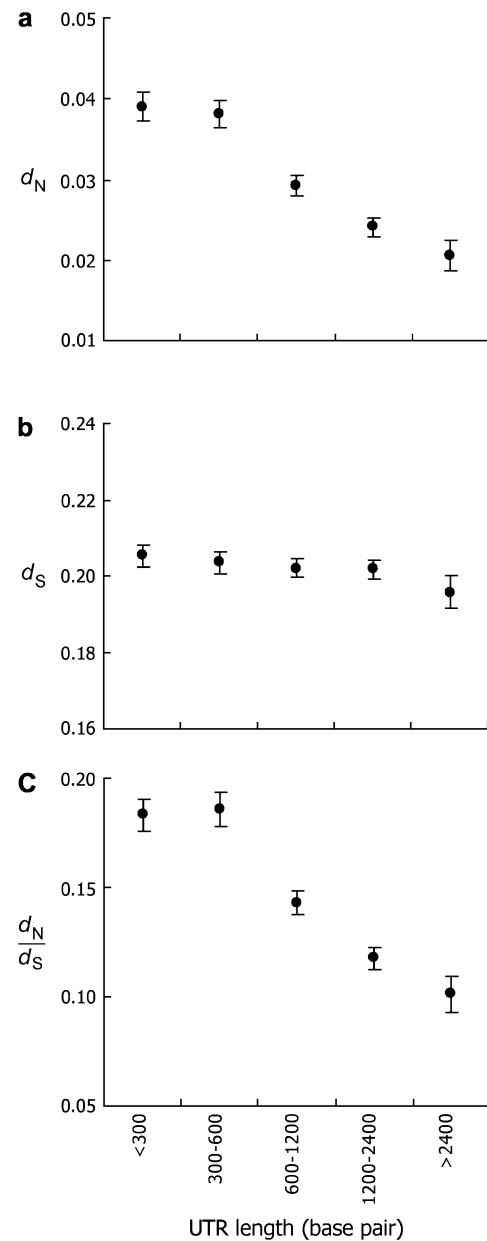


FIG. 3.—Mouse genes with longer UTRs tend to have lower  $d_N$  and  $d_N/d_S$  values. Average mouse–rat (a)  $d_N$ , (b)  $d_S$ , and (c)  $d_N/d_S$  values of each bin are shown. Error bars represent the standard error of the mean. The numbers of genes in each bin are <300: 551; 300–600: 482; 600–1,200: 673; 1,200–2,400: 622; >2,400: 247, with the total number of genes being 2,575.

Based on an analysis of 175 mouse genes, Hurst and Smith (1999) found no significant correlation between gene essentiality and  $d_N/d_S$ . Zhang and He (2005) suggested that this negative result was likely due to an insufficient sample size. Indeed, when 3,164 mouse genes are analyzed here, essential genes showed significantly lower  $d_N/d_S$  than non-essential genes. This difference remains highly significant even when we remove immune system genes and X-linked genes. Furthermore, the correlation between gene essentiality and  $d_N$  (or  $d_N/d_S$ ) is still significant after controlling for gene expression level, tissue specificity, UTR length, and intron length. We conclude that gene essentiality is an

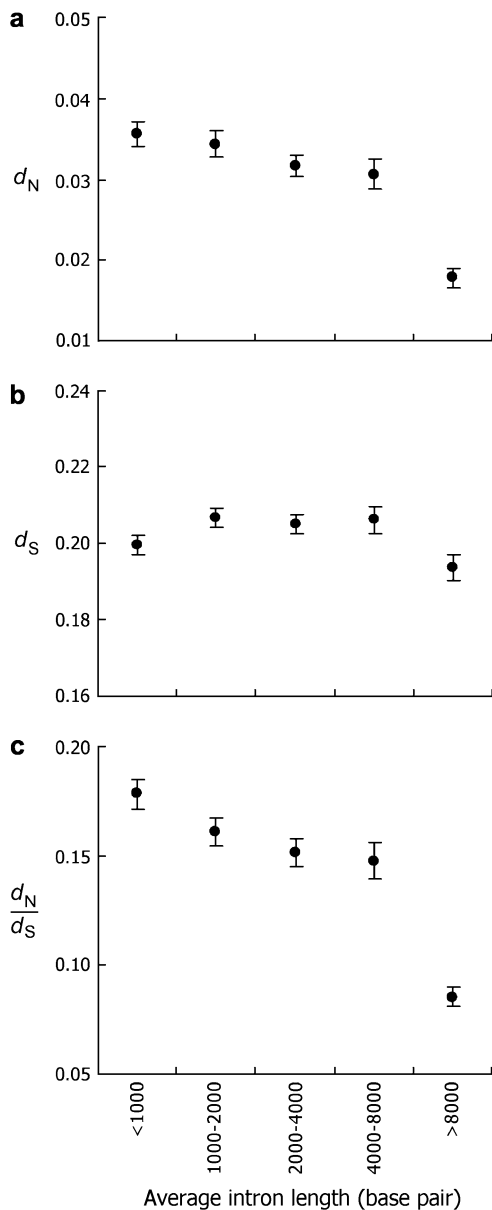


FIG. 4.—Mouse genes with larger average intron sizes tend to have lower  $d_N$  and  $d_N/d_S$  values. Average mouse–rat (a)  $d_N$ , (b)  $d_S$ , and (c)  $d_N/d_S$  values of each bin are shown. Error bars represent the standard error of the mean. The numbers of genes in each bin are <1,000: 656; 1,000–2,000: 561; 2,000–4,000: 577; 4,000–8,000: 380; >8,000: 401, with the total number of genes being 2,575.

independent determinant of the rate of mammalian protein evolution. It is interesting to note that in yeasts, the average  $d_N$  of nonessential genes is  $\sim 40\%$  higher than that of essential genes (Zhang and He 2005), a number slightly greater than that observed for mammalian genes (30%). The rank correlation coefficient between gene essentiality and  $d_N$  is  $\sim 0.2$  in yeast, also slightly greater than that in mammals (0.14). After controlling for gene expression level, the correlation coefficient becomes 0.10–0.15 in yeast and 0.13 in mammals. Note that the yeast gene knockout data used by Zhang and He (2005) contained  $>90\%$  of yeast genes, whereas the mouse gene knockout data used here contained only 15% of mouse genes. Because targeted

gene deletion in mouse requires great efforts, it is possible that researchers tend to study and report functionally important mouse genes that have human orthologs, thus reducing the variation in essentiality among the genes included in our data set. This reduction could potentially decrease the correlation coefficient between gene essentiality and  $d_N$ . But, at any rate, gene essentiality and  $d_N$  are significantly correlated in mammals. Thus, in all organisms so far examined (bacteria, yeasts, nematodes, and mammals), nonessential genes tend to evolve faster than essential genes. It is thus appropriate to conclude that the fundamental prediction of the neutral theory, that less important genes evolve faster than important genes, is universally supported by empirical data at the genomic level. However, it should be pointed out that the correlation between gene essentiality and  $d_N$ , although statistically significant, is small in magnitude. This weak correlation contrasts the strong belief of many biologists that functionally important DNA sequences evolve slowly, which is the basis of many successful bioinformatic methods such as Blast (Altschul et al. 1990) and phylogenetic footprinting (Gumucio et al. 1993). It is possible that the knockout phenotype observed in the lab only roughly reflects the amount of fitness reduction in the wild, which is expected to be a better rate determinant.

A previous study showed that human morbid genes (those known to cause diseases when mutated) evolve more slowly than nonmorbid genes (Kondrashov et al. 2004). Their analysis is not equivalent to a comparison between essential and nonessential genes because nonmorbid genes can have unidentifiable embryonic lethal phenotype or infertility phenotype when mutated. In other words, nonmorbid genes include both essential and nonessential genes, and thus, there is no clear prediction as to whether nonmorbid genes should evolve more rapidly or more slowly than morbid genes. In fact, Smith and Eyre-Walker (2003) also analyzed morbid and nonmorbid genes but obtained an opposite result.

We found that the rate of mammalian protein evolution is not, or is only weakly, correlated with the gene expression level, when gene essentiality is controlled for. In the future, it would be important to verify this finding for the entire genome as more gene knockout data become available. If our finding is generally true for mammals, it contrasts that from the yeast, where the expression level explains about a quarter ( $p^2 = \sim 0.25$ ) of the variation in  $d_N$  (Zhang and He 2005). The reduction of the correlation in mammals may be due to smaller population sizes in mammals than in yeasts because the expression level becomes a weaker selective force as the population size reduces (Ohta 1992). However, although the correlations between various rate determinants and protein evolutionary rate in mammals may be weakened by smaller population sizes, the relative importance of these rate determinants should remain unchanged. Why are the influences of gene expression level on  $d_N$  drastically different between yeast and mammals? To address this question, one has to understand why the gene expression level affects  $d_N$  in yeast. However, no widely accepted explanation exists at this time. The recently proposed translational robustness hypothesis (Drummond et al. 2005) suggests that highly expressed proteins are prone to forming misfolded protein aggregates that could be toxic or

**Table 2**  
**Partial Rank Correlation of Various Factors and  $d_N$ ,  $d_S$ , or  $d_N/d_S$** 

	$d_N$		$d_S$		$d_N/d_S$	
	$\rho$	<i>P</i> Value	$\rho$	<i>P</i> Value	$\rho$	<i>P</i> Value
Essentiality						
Essentiality   $\tau \cdot ExpLev$	-0.134	$8.12 \times 10^{-12}$	-0.047	$1.81 \times 10^{-2}$	-0.127	$8.51 \times 10^{-11}$
Essentiality   UTR · intron	-0.130	$3.71 \times 10^{-11}$	-0.046	$2.09 \times 10^{-2}$	-0.123	$3.65 \times 10^{-10}$
Expression pattern						
$\tau$   essentiality	0.165	$3.58 \times 10^{-17}$	0.092	$2.76 \times 10^{-6}$	0.149	$2.85 \times 10^{-14}$
$\tau$   UTR · intron	0.134	$8.42 \times 10^{-12}$	0.089	$6.01 \times 10^{-6}$	0.117	$2.92 \times 10^{-9}$
<i>ExpLev</i>   essentiality	-0.037	$5.97 \times 10^{-2}$	0.057	$3.62 \times 10^{-3}$	-0.048	$1.42 \times 10^{-2}$
<i>ExpLev</i>   UTR·intron	-0.046	$2.01 \times 10^{-2}$	0.057	$3.72 \times 10^{-3}$	-0.058	$3.12 \times 10^{-3}$
Gene compactness						
UTR length   $\tau \cdot ExpLev$	-0.192	$6.07 \times 10^{-23}$	-0.031	$1.16 \times 10^{-1}$	-0.194	$2.78 \times 10^{-23}$
5'-UTR length   $\tau \cdot ExpLev$	-0.131	$2.78 \times 10^{-11}$	-0.047	$1.73 \times 10^{-2}$	-0.121	$6.59 \times 10^{-10}$
3'-UTR length   $\tau \cdot ExpLev$	-0.101	$2.74 \times 10^{-7}$	-0.007	$7.16 \times 10^{-1}$	-0.106	$6.87 \times 10^{-8}$
UTR length   essentiality	-0.208	$1.51 \times 10^{-26}$	-0.039	$4.70 \times 10^{-2}$	-0.208	$1.35 \times 10^{-26}$
5'-UTR length   essentiality	-0.135	$5.07 \times 10^{-12}$	-0.047	$1.65 \times 10^{-2}$	-0.126	$1.31 \times 10^{-10}$
3'-UTR length   essentiality	-0.103	$1.63 \times 10^{-7}$	-0.009	$6.66 \times 10^{-1}$	-0.108	$4.21 \times 10^{-8}$
Intron length (avg)   $\tau \cdot ExpLev$	-0.150	$1.99 \times 10^{-14}$	0.002	$9.08 \times 10^{-1}$	-0.164	$5.51 \times 10^{-17}$
Intron length (avg)   essentiality	-0.161	$2.24 \times 10^{-16}$	-0.010	$5.96 \times 10^{-1}$	-0.161	$2.24 \times 10^{-16}$

NOTE.—Essentiality is 1 for essential genes and 0 for nonessential genes. “UTR” for UTR length and “intron” for average length per intron unless otherwise noted. The factor before “|” is the factor being examined and those after “|” are the factors being controlled for. *P* values show the probabilities of the observations under the hypothesis of no correlation. The analysis is based on 2,575 mouse genes and their rat orthologs.

pathogenic to the organism (Ellis and Pinheiro 2002). Thus, their coding regions are under intense selective pressure to maintain certain sequences that avoid misfolding in the presence of translational errors (Drummond et al. 2005). If this hypothesis is correct, our observation of no impact of expression level on  $d_N$  in mammals may be due to a lowered probability of protein aggregation in mammalian cells. It is known that a misfolded protein may aggregate, particularly when it is in high concentration (Minton 2000). The cell volume of the mouse sperm ( $61\text{--}70 \mu\text{m}^3$ ) (Brotherton 1975), the smallest mouse cell, is similar to that of a haploid yeast cell ( $\sim 70 \mu\text{m}^3$ ) (Sherman 1991). Generally speaking, other types of mammalian cells are much larger than the sperm cell (and the yeast cell). If the protein concentration (per gene) in a cell is generally lower in mammals than in yeast, the pressure of avoiding aggregation would also be lower in mammalian cells, making expression level a negligible factor in determining  $d_N$ . Nonetheless, this explanation is built on 2 assumptions, the translational robustness hypothesis and a lower protein concentration per gene in mammalian cells than in yeast cells, both of which require further scrutiny. An alternative explanation is that the gene expression level of a unicellular organism and the average gene expression level across tissues of a multicellular organism are 2 different things and are not comparable. Interestingly, when using the gene expression level estimated from the mouse expressed sequence tags (ESTs) at an embryonic stage, Subramanian and Kumar (2004) found a significant impact of gene expression level on the rate of protein evolution. Because many genes are not expressed at the embryonic stage, the biological meaning of their observation is not immediately clear. It remains to be seen whether the correlation between gene expression level and protein evolutionary rate exists only among genes having similar functions or expression patterns (as in Subramanian and Kumar’s study) but not among genes with diverse properties. Alternatively, the mi-

croarray gene expression data used in the present study may be too noisy to accurately reflect mRNA abundance compared with the EST data used by Subramanian and Kumar (2004). But, interestingly, the same microarray data revealed a strong correlation between  $\tau$  and  $d_N$ , suggesting that these data still contain a sufficient amount of expression information. We also examined the correlation between the  $d_N$  of a gene and the maximum expression level of the gene across 61 tissues surveyed. Unexpectedly, a weak positive correlation was observed ( $\rho = 0.075$ ,  $P = 1.4 \times 10^{-4}$ ). It is unclear what caused this positive correlation.

A surprising finding of the present study is that compact genes (with short UTRs and introns) tend to evolve fast (figs. 3 and 4). Although the above finding was based on genes with knockout data, essentially the same result was obtained when the entire genome was analyzed (supplementary table 4, Supplementary Material online). Previous studies showed that highly expressed genes have short introns (Castillo-Davis et al. 2002) and evolve slowly (Subramanian and Kumar 2004). Thus, one expects that genes with short introns evolve slowly. But, our observation is opposite. The reason for this unexpected observation is not entirely clear. Of course, in our analysis, gene expression level and  $d_N$  are virtually uncorrelated, and thus, the prediction that compact genes evolve slowly is invalid. Nevertheless, the observation that compact genes evolve fast is still surprising. Because UTRs and introns are noncoding regions of a gene and the majority of these sequences are more tolerant than coding regions to insertions and deletions, we consider the length variation of these noncoding sequences as a result of variation of local insertion and deletion rates (Vinogradov 2004). That is, we assume that the insertion/deletion rate ratio varies across genomic regions, making some genes more compact than others. It has been proposed that the presence and length of noncoding regions such as introns and intergenic regions can

increase the frequency of recombination between adjacent exons and genes (Comeron and Kreitman 2002). Accordingly, for 2 genes with the same functional importance, same CDS length, same number of introns, but different intron sizes, purifying selection is expected to be more efficient for the gene with bigger introns than the one with smaller introns, as the former has a higher recombination rate (per gene) than the latter. This difference results in a lower expected  $d_N$  for the gene with bigger introns, which is observed in this study. This recombination rate hypothesis also predicts a negative correlation between intron number and  $d_N/d_S$ , which is not observed (table 1). It is likely that for a given gene, intron number is much less changeable by mutation than average intron size and thus does not show the predicted correlation. Of course, recombination rate variation provides just one possible explanation of our observation; other possibilities cannot be excluded. Contrary to mammals, only 263 yeast protein-coding genes (~5%) contain intron(s). Thus, it is expected that gene compactness will not be an important factor in determining yeast protein evolution at the genomic level. However, among 86 intron-containing yeast (*S. cerevisiae*) genes that have *S. bayanus* orthologs, the average intron size and  $d_N$  are negatively correlated ( $\rho = -0.282$ ,  $P < 0.01$ ), similar to the result obtained from mammalian genes. It would be interesting to examine whether the influence of gene compactness on protein evolutionary rate is as significant as in mammals for unicellular eukaryotes with high prevalence of introns (e.g., the green algae *Chlamydomonas reinhardtii*).

In summary, we find that the relative importance of various rate determinants in mammals is gene compactness > gene essentiality  $\approx$  tissue specificity > gene expression level. This order differs substantively from that in yeasts or bacteria. For example, although the absolute magnitudes of the impact of gene essentiality are similar between the yeast and mammals, the relative impacts appear quite different because the gene expression level plays a much greater role in yeast than in mammals. It seems that the rules governing the rate of protein evolution need not be the same for all major clades of living organisms. Our results highlight the danger of applying findings from a single species, even based on a genome-wide analysis, to distantly related species and suggest reexamination of the roles of various rate determinants across a wide range of species, which is becoming feasible with the rapid advance of functional and comparative genomics.

### Supplementary Material

Supplementary tables 1–4 are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

### Acknowledgments

We thank Wendy Grus, Ondrej Podlaha, Peng Shi, and 2 anonymous reviewers for valuable comments. This work was supported by research grants from the University of Michigan and the National Institutes of Health to J.Z.

### Literature Cited

- Akashi H. 2003. Translational selection and yeast proteome evolution. *Genetics* 164:1291–303.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403–10.
- Brotherton J. 1975. The counting and sizing of spermatozoa from ten animal species using a Coulter counter. *Andrologia* 7: 169–85.
- Castillo-Davis CI, Hartl DL. 2003. Conservation, relocation and duplication in genome evolution. *Trends Genet* 19:593–7.
- Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. 2002. Selection for short introns in highly expressed genes. *Nat Genet* 31:415–8.
- Comeron JM, Kreitman M. 2002. Population, evolutionary and genomic consequences of interference selection. *Genetics* 161:389–410.
- Dayhoff MO. 1972. Atlas of protein sequence and structure. Washington, DC: National Biomedical Research Foundation.
- Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci USA* 102:14338–43.
- Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol* 23:327–37.
- Duret L, Mouchiroud D. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol Biol Evol* 17:68–74.
- Ellis RJ, Pinheiro TJ. 2002. Medicine: danger—misfolding proteins. *Nature* 416:483–4.
- Fraser HB. 2005. Modularity and evolutionary constraint on proteins. *Nat Genet* 37:351–2.
- Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002. Evolutionary rate in the protein interaction network. *Science* 296:750–2.
- Gumucio DL, Shelton DA, Bailey WJ, Slightom JL, Goodman M. 1993. Phylogenetic footprinting reveals unexpected complexity in trans factor binding upstream from the epsilon-globin gene. *Proc Natl Acad Sci USA* 90:6018–22.
- Hahn MW, Kern AD. 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol Biol Evol* 22:803–6.
- Hastings KE. 1996. Strong evolutionary conservation of broadly expressed protein isoforms in the troponin I gene family and other vertebrate gene families. *J Mol Evol* 42:631–40.
- He X, Zhang J. 2006. Toward a molecular understanding of pleiotropy. *Genetics*. Forthcoming.
- Hirsh AE, Fraser HB. 2001. Protein dispensability and rate of evolution. *Nature* 411:1046–9.
- Hubbell E, Liu WM, Mei R. 2002. Robust estimators for expression analysis. *Bioinformatics* 18:1585–92.
- Hughes AL. 1999. Adaptive evolution of genes and genomes. New York: Oxford University Press.
- Hughes AL, Nei M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335:167–70.
- Hurst LD, Smith NG. 1999. Do essential genes evolve slowly? *Curr Biol* 9:747–50.
- Jordan IK, Rogozin IB, Wolf YI, Koonin EV. 2002. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res* 12:962–8.
- Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melsopp C, Hammond M, Rocca-Serra P, Cox T, Birney E. 2004. EnsMart: a generic system for fast and flexible access to biological data. *Genome Res* 14:160–9.
- Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.

- Kimura M, Ohta T. 1974. On some principles governing molecular evolution. *Proc Natl Acad Sci USA* 71:2848–52.
- Kondrashov FA, Ogurtsov AY, Kondrashov AS. 2004. Bioinformatical assay of human gene morbidity. *Nucleic Acids Res* 32:1731–7.
- Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. 2005. Evolution of proteins and gene expression levels are coupled in *Drosophila* and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. *Mol Biol Evol* 22:1345–54.
- Li WH. 1997. Molecular evolution. Sunderland: Sinauer Associates.
- Liao BY, Zhang J. 2006a. Low rates of expression-profile divergence in highly-expressed genes and tissue-specific genes during mammalian evolution. *Mol Biol Evol* 23:1119–28.
- Liao BY, Zhang J. 2006b. Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Mol Biol Evol* 23:530–40.
- Lu J, Wu CI. 2005. Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee. *Proc Natl Acad Sci USA* 102:4063–7.
- Makino T, Gojobori T. 2006. The evolutionary rate of a protein is influenced by features of the interacting partners. *Mol Biol Evol* 23:784–9.
- Malcom CM, Wyckoff GJ, Lahn BT. 2003. Genic mutation rates in mammals: local similarity, chromosomal heterogeneity, and X-versus-autosome disparity. *Mol Biol Evol* 20:1633–41.
- Minton AP. 2000. Implications of macromolecular crowding for protein assembly. *Curr Opin Struct Biol* 10:34–9.
- Nembaware V, Crum K, Kelso J, Seoighe C. 2002. Impact of the presence of paralogs on sequence divergence in a set of mouse-human orthologs. *Genome Res* 12:1370–6.
- Ohta T. 1992. The nearly neutral theory of molecular evolution. *Annu Rev Ecol Syst* 23:263–86.
- Pal C, Papp B, Hurst LD. 2001a. Highly expressed genes in yeast evolve slowly. *Genetics* 158:927–31.
- Pal C, Papp B, Hurst LD. 2001b. Does the recombination rate affect the efficiency of purifying selection? The yeast genome provides a partial answer. *Mol Biol Evol* 18:2323–6.
- Pal C, Papp B, Hurst LD. 2003. Genomic function: rate of evolution and gene dispensability. *Nature* 421:496–7.
- Rocha EP, Danchin A. 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol Biol Evol* 21:108–16.
- Sherman F. 1991. Getting started with yeast. *Methods Enzymol* 194:3–21.
- Smith NG, Eyre-Walker A. 2003. Human disease genes: patterns and predictions. *Gene* 318:169–75.
- Su AI, Wiltshire T, Batalov S, et al. (13 co-authors). 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA* 101:6062–7.
- Subramanian S, Kumar S. 2004. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics* 168:373–81.
- Veeramachaneni V, Makalowski W, Galdzicki M, Sood R, Makalowska I. 2004. Mammalian overlapping genes: the comparative perspective. *Genome Res* 14:280–6.
- Vinogradov AE. 2004. Compactness of human housekeeping genes: selection for economy or genomic design? *Trends Genet* 20:248–53.
- Wall DP, Hirsh AE, Fraser HB, Kumm J, Gjaever G, Eisen MB, Feldman MW. 2005. Functional genomic analysis of the rates of protein evolution. *Proc Natl Acad Sci USA* 102:5483–8.
- Wang PJ, McCarrey JR, Yang F, Page DC. 2001. An abundance of X-linked genes expressed in spermatogonia. *Nat Genet* 27:422–6.
- Wilson AC, Carlson SS, White TJ. 1977. Biochemical evolution. *Annu Rev Biochem* 46:573–639.
- Winter EE, Goodstadt L, Ponting CP. 2004. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Res* 14:54–61.
- Yanai I, Benjamin H, Shmoish M, et al. (12 co-authors). 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* 21:650–9.
- Yang J, Gu Z, Li WH. 2003. Rate of protein evolution versus fitness effect of gene deletion. *Mol Biol Evol* 20:772–4.
- Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13:555–6.
- Zhang J. 2001. Protein-length distributions for the three domains of life. *Trends Genet* 16:107–9.
- Zhang J, He X. 2005. Significant impact of protein dispensability on the instantaneous rate of protein evolution. *Mol Biol Evol* 22:1147–55.
- Zhang L, Li WH. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol Biol Evol* 21:236–9.

Koichiro Tamura, Associate Editor

Accepted July 31, 2006