DOI: 10.1534/genetics.106.066027

Impact of Transcriptional Properties on Essentiality and Evolutionary Rate

Jung Kyoon Choi,* Sang Cheol Kim,† Jungmin Seo,* Sangsoo Kim[1,‡] and Jong Bhak*,[1,2]

**Korean BioInformation Center, KRIBB, Daejeon 305-806, Korea, †Department of Applied Statistics, Yonsei University, Seoul 120-749, Korea and ‡Department of Bioinformatics and Life Science, Soongsil University, Seoul 156-743, Korea*

Manuscript received September 19, 2006
Accepted for publication October 14, 2006

ABSTRACT

We characterized general transcriptional activity and variability of eukaryotic genes from global expression profiles of human, mouse, rat, fly, plants, and yeast. The variability shows a higher degree of divergence between distant species, implying that it is more closely related to phenotypic evolution, than the activity. More specifically, we show that transcriptional variability should be a true indicator of evolutionary rate. If we rule out the effect of translational selection, which seems to operate only in yeast, the apparent slow evolution of highly expressed genes should be attributed to their low variability. Meanwhile, rapidly evolving genes may acquire a high level of transcriptional variability and contribute to phenotypic variations. Essentiality also seems to be correlated with the variability, not the activity. We show that indispensable or highly interactive proteins tend to be present in high abundance to maintain a low variability. Our results challenge the current theory that highly expressed genes are essential and evolve slowly. Transcriptional variability, rather than transcriptional activity, might be a common indicator of essentiality and evolutionary rate, contributing to the correlation between the two variables.

EVOLUTION of gene expression, which has long been a subject of great interest (King and Wilson 1975), is now being studied on a genomic scale with the help of rapidly growing microarray and genome sequence data (Enard *et al.* 2002; Oleksiak *et al.* 2002; Meiklejohn *et al.* 2003; Ranz *et al.* 2003; Rifkin *et al.* 2003; Khaitovich *et al.* 2004; Denver *et al.* 2005). Of particular importance, expression level has been believed to be the best indicator of the evolutionary rate of encoded proteins. Highly expressed genes were found to evolve slowly from bacteria to mammals (Sharp 1991; Duret and Mouchiroud 2000; Pal *et al.* 2001; Herbeck *et al.* 2003; Urrutia and Hurst 2003; Subramanian and Kumar 2004; Drummond *et al.* 2005). In addition, it has recently emerged as a governing factor behind the apparent relationships between evolutionary rate and other important genomic features. Specifically, the influences of protein–protein interactions and dispensability on evolutionary rate have been disputed on the grounds that their effects may be confounded with gene expression level (Hirsh and Fraser 2001; Fraser *et al.* 2002; Bloom and Adami 2003; Pal *et al.* 2003; Bloom and Adami 2004; Fraser and Hirsh 2004; Wall *et al.* 2005). In other words, when expression level was statistically controlled, the effects decreased or disappeared. Pal *et al.* (2003) argued that essential proteins evolve more slowly only because they are highly expressed. To rule out the direct effect of essentiality on evolutionary rate, they argued on the basis of the following two hypothetical relations:

1. *Essentiality and transcriptional activity*: Each protein molecule may have the same amount of phenotypic contribution to an organism's fitness. Under this hypothesis, proteins that have more phenotypic contribution should have higher levels of active molecules in the cell. However, the validity of this hypothesis is highly questionable as genes involved in functions such as transcriptional regulation, ligand binding, and signal transduction are required only in small quantities even though they are vital for the organism. High abundance does not necessarily mean high fitness effect.
2. *Transcriptional activity and selective pressure*: Highly expressed genes may prefer translationally efficient codons, which leads to a slow rate of nucleotide sequence changes (Akashi 2001, 2003; Akashi and Gojobori 2002). Recently, using Saccharomyces as a model organism, Drummond *et al.* (2005) argued that selection may act on codon preference (for translational accuracy) and on amino acid sequence (for translational robustness) to minimize the detrimental effects of protein misfolding. Since they experience more translation events, highly expressed genes should be subject to stronger selective pressure. However, this is hardly applicable to higher eukaryotes. First, the lack of translational selection on codon preference in larger genomes is a well-known

[1] These authors contributed equally to this work.

[2] *Corresponding author:* Korean BioInformation Center, Korea Research Institute of Bioscience and Biotechnology, 52 Eoeun-dong, Yooseong-goo, Daejeon 305-333, Korea. E-mail: jong@kribb.re.kr

phenomenon (Shields *et al.* 1988; Sharp *et al.* 1995; Akashi 1997, 2001; dos Reis *et al.* 2004). Second, in a recent study (Wright *et al.* 2004), it has been shown that the number of tissues in which a gene is expressed, not the total amount of translation events across tissues, is an important determinant of evolutionary rate in Arabidopsis. We expect that this should be the case with other multicellular organisms.

Here we propose two alternative relations:

1. *Essentiality and transcriptional variability*: One can anticipate that essential proteins should have low genetic and physiological variation. To perform core functions in the cell, they are constitutively required in different individuals and physiological conditions. This leads to the expectation that transcriptional variability, which can be expressed as genetic, physiological, or random variation at different levels, may be a better indicator of essentiality than transcriptional activity. To test this hypothesis, we compared the transcriptional variability of genes associated with essential gene ontology (GO) categories to that of the other genes. For experimental validations, protein dispensability and interaction data for yeast proteins were used for a measure of essentiality.
2. *Transcriptional variability and selective pressure*: The presence of high variation among individuals in a population may indicate the action of weak purifying selection on that gene. One can envision that those genes have evolved to possess a high transcriptional variability to be expressed in specific conditions. In contrast, the genes under strong purifying selection are likely to exhibit a constant level of expression in various conditions, maintaining a limited level of variability. This explains the observation of a correlation between evolutionary rate and tissue specificity (Wright *et al.* 2004). Duplicates could be used as a good source for the quantitative measure of a correlation between expression divergence and sequence divergence (Gu 2004; Gu *et al.* 2004, 2005). Drummond *et al.* (2005) demonstrated that divergence in transcriptional activity correlated with sequence divergence between duplicates in yeast. However, we speculated that divergence in transcriptional variability would correlate better with sequence divergence in eukaryotes. We tested the second hypothesis with a genomewide analysis followed by the duplicates studies.

To provide experimental support for these hypotheses, we made use of a microarray database to extract general transcriptional properties of each gene. An average expression level across a wide range of biological conditions, including different individuals, times, tissues, disease states, environmental conditions, and so on, defined general transcriptional activity. Coefficient of variation (CV), the ratio of standard deviation (SD) over mean, was adopted as a measure of general transcriptional variability. CV has been used as a measure of stochastic fluctuation or "noise" in gene expression (Elowitz *et al.* 2002; Ozbudak *et al.* 2002; Blake *et al.* 2003; Raser and O'Shea 2004). According to the models of stochastic gene expression, noise should increase as the amount of transcript decreases (Hasty and Collins 2002; Swain *et al.* 2002). Low expression level may be coupled with high physiological or genetic variation as well as stochastic variation. Small changes in the amount of proteins that are normally expressed at a low level may have a greater impact on the cell or organism than large changes in the amount of proteins whose normal expression level is high. Therefore, it may be that there is a close correspondence between transcriptional activity and variability.

MATERIALS AND METHODS

Measuring transcriptional activity and variability: Supplemental Table 1 at http://www.genetics.org/supplemental/ shows a compendium of the multispecies microarray data obtained from the GEO database (http://www.ncbi.nlm.nih.gov/geo/). The raw Affymetrix data were used without log transformation. A flooring of the lowest 1% and ceiling of the top 1% was applied. From the microarray data sets with at least five arrays, we computed the mean and CV of expression values for each gene. To integrate different scales of microarray data sets, we scaled the mean values to $z_{ij} = (m_{ij} - \overline{m_j})/\mathrm{SD}(\widetilde{m_j})$, where m_{ij} is the mean expression level of *i*th gene in the *j*th data set, while $\overline{m_j}$ and $\mathrm{SD}(\widetilde{m_j})$ indicate the average and standard deviation of all mean values in the *j*th data set. Finally, global transcriptional properties of gene i were obtained as $\sum_{j=1}^{n} z_{ij}/n$ and $\sum_{j=1}^{n} v_{ij}/n$, namely, the mean *z*-score and mean CV across n data sets. These final estimates of transcriptional activity and variability can be downloaded at our supplemental website, http://centi.kribb.re.kr, along with all of the raw microarray data.

Evolutionary conservation index: To provide a measure of evolutionary rate on a macroevolutionary timescale, we used the HomoloGene database (http://www.ncbi.nlm.nih.gov/HomoloGene). We assigned each gene a conservation score defined as the number of its orthologs among 18 eukaryotic species. Thus, the score ranges from 1, meaning the gene is specific to one species, to 18, meaning the gene is conserved among all the 18 species. Rapidly evolving genes, including recently emerged genes (Long *et al.* 2003), will probably be conserved only in a few organisms and thus have a low conservation score. On the contrary, slowly evolving genes, such as ancient genes, will have a high conservation score. This score is expected to reflect selective pressure differently than substitution rates. As such, we used the conservation score as an inverse estimate of evolutionary rate. We refer to it as the evolutionary conservation index (ECI).

Measuring substitution rates: Orthologous relationships among human, mouse, and rat proteins were obtained from the HomoloGene database. Synonymous (d_{S}) and nonsynonymous (d_{N}) substitution rates were calculated from the alignments of coding sequences using PAML's codonml (Yang 1997). PAML's baseml was then used to estimate Jukes–Cantor distances between orthologous nucleotide sequences. A total of 12,671 human–mouse and 7154 mouse–rat ortholog pairs with observed expression pattern produced appropriate protein and nucleotide sequence alignments for PAML input. Paralogs were also identified from the HomoloGene database. We performed all-against-all comparisons between paralogs to estimate d_N, d_S, the transcriptional activity difference, and transcriptional variability difference. Transcriptional-activity divergence was measured as $|z_1 - z_2|$, where z_1 and z_2 mean the transcriptional activity of two paralogous genes. Transcriptional-variability divergence was measured as $|\mathrm{CV}_1 - \mathrm{CV}_2|$, where CV_1 and CV_2 mean the transcriptional variability of two paralogous genes.

Principal-component analysis-based regression: Although similar in principle, the two representative principal-component analysis (PCA)-based regression methods have different purposes. Principal-component regression (PCR) determines the linear combinations of the predictors that explain most of the variation in these predictors, ignoring the response variable. Partial least-squares regression (PLS), in contrast, finds the linear combinations that best explain the response, yielding different results according to the type of the response (in this case, ECI, d_N, d_S, the fitness effect, and protein interaction number). Although practically it produces similar results to PLS, PCR is not designed to determine the influence of the predictors on the response. In fact, PCR failed to distinguish the contribution of the predictors (transcriptional activity and variability) in many cases in our analysis. The R package "pls" was used to perform PLS. We scaled the predictors to zero mean and unit variance before conducting PLS.

Functional category analysis: For each GO category with >10 genes, we computed a normalized transcriptional variability (*z*-score) from the CVs of all members of that category. The *z*-score was defined as $z = (\overline{v_{\mathrm{GO}}} - \overline{v_{\mathrm{all}}})/(\mathrm{SD}(\widetilde{v_{\mathrm{GO}}})/\sqrt{n_{\mathrm{GO}}})$, where $\overline{v_{\mathrm{GO}}}$ is the average CV for the genes in the GO category, $\overline{v_{\mathrm{all}}}$ is the average CV for all genes in the species, $\mathrm{SD}(\widetilde{v_{\mathrm{GO}}})$ is the standard deviation for the genes in the category, and n_{GO} is the number of the genes in the category. The *z*-scores are listed for each species in the supporting information available at http://centi.kribb.re.kr. For each category, we obtained a specieswise *z*-score as $\sum_{k=1}^{n} z_k/n$, where z_k indicates the *z*-score in the *k*th species when $n \geq 3$. The specieswise *z*-scores are given in supplemental Table 4 at http://www.genetics.org/supplemental/. The significant categories with more than a total of 100 genes across the species were selected for presentation in supplemental Tables 2 and 3 at http://www.genetics.org/supplemental/.

RESULTS AND DISCUSSION

Evolution of general transcriptional properties: We collected Affymetrix microarray data for the species for which a considerable amount of expression profiles have been produced (supplemental Table 1 at http://www.genetics.org/supplemental/). For example, 177 different data sets containing 16,446 HomoloGene entries in 4136 arrays were gathered as a human expression profile. The mean and CV of expression values were obtained for each gene. The mean value was scaled to a *z*-score. The average *z*-score and CV were computed across the data sets within a species to represent general transcriptional activity and variability, respectively.

We needed to check if the expression profiles used in the analysis contained a sufficient amount of data to convey information on the universal transcriptional properties of each gene. To test this, we made comparisons of the *z*-score and CV among the species on the assumption that the transcriptional properties should be evolutionarily conserved. Since human, mouse, and rat had much more microarray data sets than the other species (supplemental Table 1 at http://www.genetics.org/supplemental/), we randomly selected 20 data sets for each species to remove the sample-size effect. For HomoloGene entries whose transcriptional properties were observed in more than one species, we calculated the correlation coefficient for every species pair and found striking correlations in most cases (Table 1). The transcriptional activity is well conserved from yeast to human. Whereas maintaining a high level of similarity among human, mouse, and rat, the variability shows some degree of divergence between distant species. Using the whole data for human, mouse, and rat resulted in the same patterns. These findings suggest that our measures of transcriptional activity and variability reflect inherent biological features that are highly conserved among orthologs, while remaining free of sampling problems. Moreover, the higher level of divergence in the variability over that measured in the activity implies that transcriptional variability might play a more important role in the evolution of phenotypic variations.

TABLE 1

Correlation of transcriptional properties among orthologs

	Human	Mouse	Rat	Fly	Plant	Yeast
Human		0.460	0.428	0.213	0.246	0.165
Mouse	0.469		0.454	0.240	0.263	0.190
Rat	0.443	0.507		0.256	0.258	0.237
Fly	0.352	0.380	0.404		0.193	0.164
Plant	0.318	0.395	0.387	0.418		0.248
Yeast	0.385	0.446	0.467	0.511	0.511	

For 31,046 HomoloGene entries whose expression pattern was observed in one or more species, the specieswise correlations of transcriptional activity (below the diagonal) and transcriptional variability (above the diagonal) were computed. To rule out the effect of sample-size bias, we randomly selected 20 data sets for human, mouse, and rat. However, the use of the selected data did not show differences from that of the whole data, underscoring that our measures reflect inherent biological features without regard to sample size. Spearman's rank correlation coefficient was used.

Transcriptional activity *vs.* transcriptional variability: To address our main question, we essentially needed to measure the correlations of these transcriptional properties with various response variables, namely estimates of essentiality and evolutionary rate. Spearman's rank correlation coefficient was used since it is robust to outliers and able to properly handle different scales of various estimates. We denote the correlation of transcriptional activity as R_a and that of transcriptional variability as R_v. A primary concern during the analysis was the association between transcriptional activity and

TABLE 2

Comparative analysis of the influences of transcriptional activity and variability on essentiality in yeast and mouse

Species	R_a	R_v	$R_{a\mid v}$	$R_{v\mid a}$	C_a	C_v	PC_a	PC_v
				Fitness effect				
Yeast	0.152	−0.280	0.057	$-\underline{0.244}$	*−0.698*	$-\underline{15.966}$***	21.45	$\underline{78.55}$
				Lethality				
Mouse	0.103	−0.153	0.015	$-\underline{0.115}$	0.621	$-\underline{6.972}$***	36.88	$\underline{63.12}$
				Protein–protein interaction				
Yeast	0.123	−0.215	0.049	$-\underline{0.184}$	*−0.132*	$-\underline{10.786}$***	18.30	$\underline{81.70}$

Transcriptional activity (a) and variability (v) are the predictor variables. Fitness effect, lethality, and protein–protein interaction are the response variables. R_a and R_v denote Spearman's rank correlation. $R_{a|v}$ means R_a controlling for v and $R_{v|a}$ means R_v controlling for a. C_a and C_v are the *t*-values of regression coefficients from multivariate regression analysis. The significance of the *t*-values is represented in Tables 2–5 as $***P < 10^{-9}$, $**P < 10^{-6}$, and $*P < 10^{-3}$. PC_a and PC_v mean the percentage of contribution of the predictors to the first principal component estimated from PCA-based regression. The statistics are underlined when they are significantly greater for one predictor than for the other. Italics are used where the statistics are positive (or negative) when negative (or positive) values are expected.

variability. As mentioned in the Introduction, there seems to be a correspondence between them. We observed the pattern as expected in all the species (data not shown). This led to the use of a partial correlation with which we can measure the correlation of transcriptional activity (or variability) with the response variables when controlling for transcriptional variability (or activity). We denote it as $R_{a|v}$ (or $R_{v|a}$). We also employed multivariate regression to compare the influences of the two predictors on the response variables, as estimated by the regression coefficient or slope (C_a and C_v).

According to Drummond *et al.* (2006), noisy variables may result in spurious partial correlations. In this regard, they suggest the use of a PCA-based regression approach. However, if one of the original correlations, R_a or R_v, is found to be much higher than the other, the noise problem will not change the conclusion based on the partial correlations. Therefore, we can simply ask if $|R_a| > |R_v|$ (or $|R_v| > |R_a|$) when $R_{a|v}$ (or $R_{v|a}$) is significant. However, we also made use of a PCA-based regression method as suggested. We present the percentage of contributions of transcriptional activity and variability to the first principal component as PC_a and PC_v in the following sections.

First hypothesis—essentiality and transcriptional variability: To estimate essentiality, we adopted data from yeast and mouse deletion experiments. Growth rates of yeast deletion strains were measured by an array-based method (Giaever *et al.* 2004). Using this data set, Wall *et al.* (2005) recently reported a positive relationship between dispensability and evolutionary rate. As an inverse measure of essentiality, we used the fitness effect, $f(i)$, as $1 - g(i)/g(\max)$, where $g(i)$ is the growth rate of the strain with gene *i* deleted, and $g(\max)$ is the maximal growth rate (Hirsh and Fraser 2001). Mouse genes subject to deletion experiments were obtained from the Mouse Genome Informatics (MGI) database (http://www.informatics.jax.org). We selected genes with the knockout phenotype of lethality (MP:0005373, "lethality postnatal"; MP:0005374, "lethality embryonic/perinatal"). The phenotype database included 1427 essential genes (lethality = 1) and 1956 nonessential genes (lethality = 0) according to our criteria. R_a and R_v for the fitness effect and lethality were estimated (Table 2). We observed significant partial correlations of transcriptional variability ($R_{v|a} = -0.244$ and -0.115) while $|R_v| > |R_a|$. On the contrary, transcriptional activity showed weak correlations ($R_{a|v} = 0.057$ and 0.015). The regression methods confirmed the disproportionate contribution of transcriptional variability to fitness effect and lethality. Hubs in protein networks are known to be essential, which prompted us to use the number of protein interactions as an estimate of essentiality (He and Zhang 2006). The same trend was found in our data (Table 2), which is in good agreement with the previously reported negative relationship between genetic variation in gene expression and the number of protein interactions (Lemos *et al.* 2004). It was also shown that this relationship was not confounded by gene expression level.

To extend this conclusion to other taxa, a functional category analysis was carried out. For each GO category, the average CV of the genes in the category was obtained, normalized to a *z*-score in each species, and converted to a specieswise *z*-score (see materials and methods). Significantly negative *z*-scores indicate that the genes in the category exhibit a low variability relative to the average of all genes in the species. Supplemental Table 2 at http://www.genetics.org/supplemental/ lists the GO terms with significantly low *z*-scores while supplemental Table 3 at http://www.genetics.org/supplemental/ shows those with significantly high *z*-scores (the full list of the

TABLE 3

Comparative analysis of the influences of transcriptional activity and variability on the evolutionary conservation index

Species	R_{a}	R_{v}	$R_{\mathrm{a}\mid\mathrm{v}}$	$R_{\mathrm{v}\mid\mathrm{a}}$	C_{a}	C_{v}	PC$_{\mathrm{a}}$	PC$_{\mathrm{v}}$
				ECI				
Human	0.207	−0.242	0.074	−0.146	*−0.051*	−29.981***	33.13	66.88
Mouse	0.245	−0.312	0.053	−0.207	2.634	−32.972***	39.58	60.42
Rat	0.233	−0.273	0.080	−0.166	3.739*	−21.267***	39.85	60.15
Fly	0.259	−0.296	0.099	−0.178	0.383	−23.521***	37.36	62.64
Plant	0.260	−0.348	0.046	−0.244	5.543**	−24.293***	40.13	59.87
Yeast	0.294	−0.196	0.248	−0.111	12.364***	−9.661***	53.78	46.22

Transcriptional activity (a) and variability (v) are the predictor variables and ECI is the response variable. See the Table 2 legend for details.

specieswise *z*-scores is given in supplemental Table 4, and the species-specific *z*-scores are available at our supplemental website, http://centi.kribb.re.kr). We find that the GO categories in supplemental Table 2 are biased toward essential cellular processes such as transcriptional and translational regulation, protein folding and transport, protein catabolism, protein complexation, RNA processing, etc. In terms of the cellular component, the genes located in the nucleus, cytoplasm, mitochondrion, and Golgi apparatus show low transcriptional variability. In contrast, supplemental Table 3 is enriched for the GO terms associated with extracellular communication such as immune response, cell–cell signaling, cell adhesion, surface receptors, growth factor activity, hormone activity, chemotaxis, etc. In the same context, the genes located in the extracellular space and plasma membrane show high variability. Notably, these findings are strikingly consistent with the results of a recent study by Chuang and Li (2004). Most of the GO categories reported to have low substitution rates are found in our list of the GO categories with less variable genes, and vice versa. Likewise, there is a remarkable overlap between the GO categories with high substitution rates and those with highly variable genes. These findings are suggestive of correspondence between transcriptional variability and substitution rate.

Evolutionary conservation on a macroevolutionary timescale: As this study spans a long evolutionary time from yeast to human, we needed to estimate evolutionary rates differently from substitution rates between paired nucleotide sequences. One solution was to count the number of genomes in which the gene is present. This approach was successfully used in the previous studies about the relationships between evolutionary rate and protein–protein interactions or protein dispensability (Jordan *et al.* 2002, 2003; Wuchty *et al.* 2003).

The ECI produced results that are consistent with conventional substitution rates. Substitution rates between orthologs cannot be reliably used for distantly related species. Therefore, we calculated them between mammalian orthologs and compared them to the ECI data. Strong inverse correlations were found: Spearman's $R = -0.362$–∼-0.420 (P-values $\ll 10^{-16}$). Moreover, the correlations of ECI with protein–protein interactions (Spearman's $R = 0.224$) and dispensability (Spearman's $R = -0.270$) were compatible with the reported correlations of substitution rates with protein–protein interactions (Spearman's $R = -0.21$, Fraser *et al.* 2002) and dispensability (Spearman's $R = 0.230$, Wall *et al.* 2005). However, using different reference species for the computation of substitution rates has fueled arguments about the association of protein–protein interactions and dispensability with evolutionary rate (Jordan *et al.* 2003; Pal *et al.* 2003; Wall *et al.* 2005). In contrast, the ECI is expected to offer a global and unified measure of evolutionary rate for diverse organisms spanning a long evolutionary time period.

Second hypothesis—transcriptional variability and selective pressure: Tables 3–5 show the statistical analysis results regarding the relationship between evolutionary rate and transcriptional properties. First, as a genomewide analysis, we investigated the influence of transcriptional activity and variability on ECI. As shown in Table 3, the correlation, partial correlation, multivariate regression, and PCA-based regression analyses all indicate that the influence of variability is much stronger than that of activity. This tendency holds true when we use traditional substitution rates between mammalian orthologs (Table 4). The only exception occurred in yeast and is discussed later. Using ECI as the response variable, all the different statistical techniques showed the same pattern, in favor of transcriptional activity in yeast and variability in the other species. Moreover, the levels of the statistics are comparable among various species (*e.g.*, $-0.1 < R_{\mathrm{v}\mid\mathrm{a}} < -0.25$, $-21 < C_{\mathrm{v}} < -33$).

Next, we extracted paralog information from the HomoloGene database. Accelerated evolution of duplicate gene expression has been shown in yeast and fruit fly (Gu *et al.* 2004). Here, we compared the transcriptional

TABLE 4

Comparative analysis of the influences of transcriptional activity and variability on substitution rates between mammalian orthologs

Species pair	R_a	R_v	$R_{a\|v}$	$R_{v\|a}$	C_a	C_v	PC_a	PC_v
				d_N				
Human	−0.186	0.201	−0.079	0.111	−6.341***	18.425***	41.52	58.48
Mouse	−0.222	0.252	−0.079	0.146	−7.389***	21.011***	44.01	55.99
Mouse	−0.210	0.251	−0.065	0.155	−6.380***	15.110***	45.21	54.79
Rat	−0.162	0.208	−0.042	0.139	−5.202**	12.695***	43.63	56.37
				d_S				
Human	−0.093	0.165	*0.012*	0.137	*1.034*	16.001***	30.83	69.17
Mouse	−0.075	0.128	*0.010*	0.104	*0.628*	11.695***	36.30	63.70
Mouse	−0.075	0.133	*0.014*	0.111	*0.028*	8.676***	38.27	61.73
Rat	−0.093	0.132	−0.014	0.095	−2.492	8.157***	41.93	58.07
				Jukes–Cantor distance				
Human	−0.179	0.201	−0.070	0.116	−4.541*	19.266***	39.35	60.65
Mouse	−0.196	0.234	−0.059	0.143	−4.403*	20.752***	41.97	58.03
Mouse	−0.180	0.230	−0.043	0.152	−3.470*	14.960***	42.63	57.37
Rat	−0.148	0.188	−0.039	0.124	−3.738*	12.169***	42.10	57.90

Substitution rates such as d_N, d_S, and Jukes–Cantor distance were calculated as described in MATERIALS AND METHODS. Transcriptional activity (a) and variability (v) are the predictor variables and the substitution rates are the response variables. See the Table 2 legend for details.

variability of duplicates with that of singletons in all the species for which we obtained data. As shown in supplemental Figure 1 at http://www.genetics.org/supplemental/, the distributions shifted toward high CV values with gene duplications, indicating increased transcriptional variability among duplicates. We also observed that overall CV values are relatively lower in plants and in yeast as compared to higher eukaryotes. We next asked whether sequence divergence, measured by d_N and d_S, is better explained by transcriptional-activity divergence ($|z_1 - z_2|$, where z_1 and z_2 mean the transcriptional activity of two paralogous genes) or by transcriptional-variability divergence ($|CV_1 - CV_2|$, where CV_1 and CV_2 mean the transcriptional variability

TABLE 5

Comparative analysis of the influences of transcriptional-activity divergence and transcriptional-variability divergence on sequence divergence between paralogous genes

Species	R_{ad}	R_{vd}	$R_{ad\|vd}$	$R_{vd\|ad}$	C_{ad}	C_{vd}	PC_{ad}	PC_{vd}
				d_N				
Human	0.126	0.258	*−0.139*	0.264	*−2.443*	3.322*	20.73	79.27
Mouse	0.140	0.230	*−0.119*	0.219	*−4.489*	7.131***	23.19	76.81
Rat	0.174	0.157	0.093	0.053	*−1.466*	19.670***	13.80	86.20
Fly	0.057	0.152	*−0.358*	0.382	*−3.333*	10.404***	4.14	95.86
Plant	0.348	0.351	0.017	0.048	1.338	3.939*	34.71	65.29
Yeast	0.132	0.051	0.124	*−0.023*	*−0.037*	0.907	44.21	55.79
				d_S				
Human	0.176	0.298	*−0.108*	0.265	*−0.749*	3.140	1.55	98.45
Mouse	0.218	0.250	0.005	0.126	*−1.721*	3.274	29.65	70.35
Rat	0.164	0.179	0.059	0.093	0.764	18.286***	19.57	80.43
Fly	0.288	0.354	*−0.228*	0.310	0.812	6.031**	25.86	74.14
Plant	0.270	0.268	0.036	0.013	2.920	2.338	53.52	46.48
Yeast	0.141	0.097	0.105	0.026	0.721	0.968	49.22	50.78

Transcriptional-activity divergence (ad) and transcriptional-variability divergence (vd) are the predictor variables. Synonymous substitution rate (d_S) and nonsynonymous substitution rate (d_N) are the response variables. Transcriptional-activity divergence was measured as $|z_1 - z_2|$, where z_1 and z_2 mean the transcriptional activity of two paralogous genes. Transcriptional-variability divergence was measured as $|CV_1 - CV_2|$, where CV_1 and CV_2 mean the transcriptional variability of two paralogous genes. See the Table 2 legend for details.

of two paralogous genes). For human, mouse, and fly, all the results are supportive of our hypothesis (Table 5). In the case of rat, although the correlation and partial correlation analysis failed to detect a specific trend, the regression analysis results strongly support our hypothesis. The signals are relatively weak in plants and yeast; nonetheless, in agreement with the ECI results, the partial correlations indicate that the effect of transcriptional activity is stronger in yeast.

Do highly expressed genes evolve slowly? We have shown that contrary to the consensus that transcriptionally active genes evolve slowly, evolution shapes transcriptional variability rather than transcriptional activity. Selective pressure seems to act primarily on transcriptional variability. If this is so, how can we explain the apparent correlation between transcriptional activity and evolutionary rate observed in the previous studies?

First, as described in the Introduction, translational selection has mainly explained the correlation (Akashi 2001, 2003; Akashi and Gojobori 2002; Drummond *et al.* 2005). Tables 3 and 5 show that evolutionary rate is still correlated with transcriptional activity in yeast, even after controlling for transcriptional variability. Therefore, if we rule out the effect of translational selection such as in higher eukaryotes (Shields *et al.* 1988; Sharp *et al.* 1995; Akashi 1997, 2001; dos Reis *et al.* 2004; Wright *et al.* 2004), we can conclude that invariable genes evolve slowly.

Second, from the inverse relationship between the mean and CV, we can speculate that as transcriptional activity increases, relative transcriptional variability will tend to decrease. In biological terms, expression changes of abundant proteins are likely to have smaller effects on the cell than those of scanty proteins. High expression levels may confer tolerance to a fluctuation in the amount, while low expression levels may enable delicate transcriptional regulation. From a stochastic perspective, a high level of transcriptional activity leads to the reduction of random noise, which means a reduced variability. This explanation can be also applied to the correlation of expression level and essentiality. Indeed, the production of essential proteins was shown to involve lower levels of noise (Fraser *et al.* 2004). These speculations suggest that a high abundance of slowly evolving or essential proteins may be evolutionarily favored to maintain a low variability in expression.

Taken together, the two aspects, the action of translational selection in yeast and the correspondence between transcriptional activity and variability, may explain the apparent correlation between transcriptional activity and evolutionary rate in the previous studies.

Conclusion: In this study, we characterized general transcriptional activity and variability of eukaryotic genes from global expression profiles of various species spanning a long evolutionary time period. While the transcriptional properties were shown to be remarkably conserved during the evolutionary processes, the variability showed a higher degree of divergence between distant species. Transcriptional variability might be related to phenotypic variations and thus be more subject to selective pressure. Indeed, we showed that transcriptional variability should be a true indicator of evolutionary rate. If we rule out the effect of translational selection, which seems to operate only in yeast, the apparent slow evolution of highly expressed genes should be attributed to their low variability. Selective forces may enable phenotypic variations to evolve mainly by shaping transcriptional variability. Furthermore, we suggest that a high abundance of essential proteins may be favored to maintain a low variability in their amount. Transcriptional variability, rather than transcriptional activity, might be a common indicator of essentiality and evolutionary rate, contributing to the correlation between the two variables.

We are grateful to the members of the Integrative Evolutionary Analysis Team, ArrayPort Team, and Core Team at Korean BioInformation Center for producing the data and maintaining the computer systems we used in the analysis. This work was supported by the Korean Ministry of Science and Technology under grant no. M10407010001-04N0701-00110.

LITERATURE CITED

Akashi, H., 1997 Codon bias evolution in *Drosophila.* Population genetics of mutation-selection drift. Gene **205:** 269–278.

Akashi, H., 2001 Gene expression and molecular evolution. Curr. Opin. Genet. Dev. **11:** 660–666.

Akashi, H., 2003 Translational selection and yeast proteome evolution. Genetics **164:** 1291–1303.

Akashi, H., and T. Gojobori, 2002 Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis.* Proc. Natl. Acad. Sci. USA **99:** 3695–3700.

Blake, W. J., M. Kaern, C. R. Cantor and J. J. Collins, 2003 Noise in gene expression. Nature **422:** 633–637.

Bloom, J. D., and C. Adami, 2003 Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets. BMC Evol. Biol. **3:** 21.

Bloom, J. D., and C. Adami, 2004 Evolutionary rate depends on number of protein-protein interactions independently of gene expression level: response. BMC Evol. Biol. **4:** 14.

Chuang, J. H., and H. Li, 2004 Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome. PLoS Biol. **2:** e29.

Denver, D. R., K. Morris, J. T. Streelman, S. K. Kim, M. Lynch *et al.*, 2005 The transcriptional consequences of mutation and natural selection in *Caenohabditis elegans.* Nat. Genet. **37:** 544–548.

dos Reis, M., R. Savva and L. Wernisch, 2004 Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. **32:** 5036–5044.

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.

Elowitz, M. B., A. J. Levine, E. D. Siggia and P. S. Swain, 2002 Stochastic gene expression in a single cell. Science **297:** 1183–1186.

Enard, W., P. Khaitovich, J. Klose, S. Zollner, F. Heissig *et al.*, 2002 Intra- and interspecific variation in primate gene expression patterns. Science **296:** 340–343.

Fraser, H. B., and A. E. Hirsh, 2004 Evolutionary rate depends on number of protein-protein interactions independently of gene expression level. BMC Evol. Biol. **4:** 13.

Fraser, H. B., A. E. Hirsh, L. M. Steinmetz, C. Scharfe and M. W. Feldman, 2002 Evolutionary rate in the protein interaction network. Science **296:** 750–752.

Fraser, H. B., A. E. Hirsh, G. Giaever, J. Kumm and M. B. Eisen, 2004 Noise minimization in eukaryotic gene expression. PLoS Biol. **2:** e137.

Giaever, G., P. Flaherty, J. Kumm, M. Proctor, C. Nislow *et al.*, 2004 Chemogenomic profiling: identifying the functional interactions of small molecules in yeast. Proc. Natl. Acad. Sci. USA **101:** 793–798.

Gu, X., 2004 Statistical framework for phylogenomic analysis of gene family expression profiles. Genetics **167:** 531–542.

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., 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.

Hasty, J., and J. J. Collins, 2002 Translating the noise. Nat. Genet. **31:** 13–14.

He, X., and J. Zhang, 2006 Why do hubs tend to be essential in protein networks? PLoS Genet. **2:** e88.

Herbeck, J. T., D. P. Wall and J. J. Wernegreen, 2003 Gene expression level influences amino acid usage, but not codon usage, in the tsetse fly endosymbiont *Wigglesworthia*. Microbiology **149:** 2585–2596.

Hirsh, A. E., and H. B. Fraser, 2001 Protein dispensability and rate of evolution. Nature **411:** 1046–1048.

Jordan, I. K., I. B. Rogozin, Y. I. Wolf and E. V. Koonin, 2002 Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res. **12:** 962–968.

Jordan, I. K., Y. I. Wolf and E. V. Koonin, 2003 No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. BMC Evol. Biol. **3:** 1.

Khaitovich, P., G. Weiss, M. Lachmann, I. Hellmann, W. Enard *et al.*, 2004 A neutral model of transcriptome evolution. PLoS Biol. **2:** e132.

King, M. C., and A. C. Wilson, 1975 Evolution at two levels: molecular similarities and biological differences between humans and chimpanzees. Science **188:** 107–116.

Lemos, B., C. D. Meiklejohn and D. L. Hartl, 2004 Regulatory evolution across the protein interaction network. Nat. Genet. **36:** 1059–1060.

Long, M., E. Betran, K. Thornton and W. Wang, 2003 The origin of new genes: glimpses from the young and old. Nat. Rev. Genet. **4:** 865–875.

Meiklejohn, C. D., J. Parsch, J. M. Ranz and D. L. Hartl, 2003 Rapid evolution of male-based gene expression in *Drosophila*. Proc. Natl. Acad. Sci. USA **100:** 9894–9899.

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.

Ozbudak, E. M., M. Thattai, I. Kurtser, A. D. Grossman and A. van Oudenaarden, 2002 Regulation of noise in the expression of a single gene. Nat. Genet. **31:** 69–73.

Pal, C., B. Papp and L. D. Hurst, 2001 Highly expressed genes in yeast evolve slowly. Genetics **158:** 927–931.

Pal, C., B. Papp and L. D. Hurst, 2003 Rate of evolution and gene dispensability. Nature **421:** 496–497.

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.

Raser, J. M., and E. K. O'Shea, 2004 Control of stochasticity in eukaryotic gene expression. Science **304:** 1811–1814.

Rifkin, S. A., J. Kim and K. P. White, 2003 Evolution of gene expression in the *Drosophila melanogaster* subgroup. Nat. Genet. **33:** 138–144.

Sharp, P. M., 1991 Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium*: codon usage, map position, and concerted evolution. J. Mol. Evol. **33:** 23–33.

Sharp, P. M., M. Averof, A. T. Lloyd, G. Matassi and J. F. Peden, 1995 DNA sequence evolution: the sounds of silence. Philos. Trans. R. Soc. Lond. Ser. B **349:** 241–247.

Shields, D. C., P. M. Sharp, D. G. Higgins and F. Wright, 1988 "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. **5:** 704–716.

Subramanian, S., and S. Kumar, 2004 Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics **168:** 373–381.

Swain, P. S., M. B. Elowitz and E. D. Siggia, 2002 Intrinsic and extrinsic contributions to stochasticity in gene expression. Proc. Natl. Acad. Sci. USA **99:** 12795–12800.

Urrutia, A. O., and L. D. Hurst, 2003 The signature of selection mediated by expression on human genes. Genome Res. **13:** 2260–2264.

Wall, D. P., A. E. Hirsh, H. B. Fraser, J. Kumm, G. Giaever *et al.*, 2005 Functional genomic analysis of the rates of protein evolution. Proc. Natl. Acad. Sci. USA **102:** 5483–5488.

Wright, S. I., C. B. K. Yau, M. Looseley and B. C. Meyers, 2004 Effects of gene expression on molecular evolution in *Arabidopsis thaliana* and *Arabidopsis lyrata*. Mol. Biol. Evol. **21:** 1719–1726.

Wuchty, S., Z. N. Oltvai and A.-L. Barabasi, 2003 Evolutionary conservation of motif constituents in the yeast protein interaction network. Nat. Genet. **35:** 176–179.

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

Communicating editor: P. J. Oefner