Tissue-driven hypothesis of genomic evolution and sequence-expression correlations

Xun Gu*†‡ and Zhixi Su*

*School of Life Sciences, Institutes of Biomedical Sciences, Center for Evolutionary Biology, Fudan University, Shanghai 200433, China; and †Department of Genetics, Development, and Cell Biology, Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA 50011

Communicated by Jiazhen Tan, Fudan University, Shanghai, China, December 18, 2006 (received for review September 30, 2006)

To maintain normal physiological functions, different tissues may have different developmental constraints on expressed genes. Consequently, the evolutionary tolerance for genomic evolution varies among tissues. Here, we formulate this argument as a ''tissue-driven hypothesis'' based on the stabilizing selection model. Moreover, several predicted genomic correlations are tested by the human–mouse microarray data. Our results are as follows. First, between the human and mouse, we have elaborated the among-tissue covariation between tissue expression distance (E_{ti}) and tissue sequence distance (D_{ti}). This highly significant $E_{ti} - D_{ti}$ correlation emerges when the expression divergence and protein sequence divergence are under the same tissue constraints. Second, the tissue-driven hypothesis further explains the observed significant correlation between the tissue expression distance (between the human and mouse) and the duplicate tissue distance (T_{dup}) between human (or mouse) paralogous genes. In other words, between-duplicate and interspecies expression divergences covary among tissues. Third, for genes with the same expression broadness, we found that genes expressed in more stringent tissues (e.g., neurorelated) generally tend to evolve more slowly than those in more relaxed tissues (e.g., hormone-related). We conclude that tissue factors should be considered as an important component in shaping the pattern of genomic evolution and correlations.

expression divergence | tissue expression | mammalian genomics | gene duplications

Understanding the underlying evolutionary mechanism is fundamental for investigating the emergence of genome complexity (1–2). It remains highly controversial as to what factors could determine the evolutionary rate of expression and sequence divergence (3–15). An important issue is the role of tissue-specific factors in genomic evolution. Several studies have suggested that tissue-specific constraints may generate among-tissue variation of expression divergence between human and chimpanzee (3, 4, 6, 8), or between human and mouse (16). Moreover, it has been found (17, 18) that the rate of expression divergence may be negatively associated to the broadness of tissue expression of the gene. Interestingly, Rifkin *et al.* (19) reported that, relative to the prediction of strict neutral model (1), the natural expression variation in the *Drosophila* population was constrained. However, there are many debates among authors about measures for expression divergence and tissue specificity, biological/statistical issues for expression-sequence correlation, and methods for multitissue analysis (see refs. 14 and 15 for recent reviews). As a working hypothesis, it seems generally accepted that correlated genomic evolution is mainly driven by various stabilizing (or purifying) selections including at the tissue level (6, 7, 19, 20). This view does not exclude the role of natural selection, which may occur in some lineages for some genes that perform specified functions. A good example is from the analyses of refs 3 and 4, implying adaptive expression shifts of some genes in the human brain. However, the opposite view remains (9, 21), arguing that expression divergence was mainly driven by natural selection.

We have recognized that, without developing an explicit evolutionary model that can provide a common ground for predicting and testing by coherent data analyses, it is difficult to have a comprehensive understanding of these issues. In this article, we develop a stochastic model for genomic evolution under the principle of stabilizing selection and formulate the tissue-driven hypothesis by postulating that stabilizing selections for both expression and sequence divergences may be affected simultaneously by the common factors of tissues in which the genes is expressed. Facilitated by substantial multispecies microarrays (16), we test several predicted genomic correlations from the tissue-driven hypothesis. Finally, we discuss the evolutionary scenario of genomic correlations, demonstrating that accumulated tissue constraints may shape the correlated pattern of sequence and expression evolution.

The Model

Expression Divergence Under Stabilizing Model. We modified the stabilizing selection model (22) of quantitative characters to describe the tissue-specific constraint on expression divergence. For a gene expressed in a certain tissue (*ti*), the stabilizing selection on the expression level *x* follows a Gaussian fitness function

$$f_{ti}(x) = e^{-w_{ti}(x-\theta_e)2}, \qquad \textbf{[1]}$$

where θ_e is the optimal value of expression level, w_{ti} is the coefficient for stabilizing selection on gene expression in tissue *ti*; a large w_{ti} means a strong selection pressure, and vice versa (Fig. 1). Under the stabilizing model of Eq. **1**, we have shown that the expression divergence follows an Ornstein–Uhlenback (OU) process (23). The stochastic OU process is characterized by the infinitesimal mean $-\beta_0(x - \theta_e)$ and variance $\varepsilon^2/2N_e$, where ε^2 is the mutational variance, N_e is the effective population size, and $\beta_0 = w_{ti\varepsilon}{}^2$ measures the direct force against the deviation from the optimum. Given the initial expression value x_0, the OU model claims that $x(t)$ follows a normal distribution with the mean and variance given by

$$E[x(t)|x_0] = e^{-\beta t}x_0 + (1 - e^{-\beta t})\theta_e \quad \text{and} \qquad \textbf{[2]}$$

$$V[x(t)|x_0] = \frac{\epsilon^2(1 - e^{-2\beta t})}{2\beta},$$

respectively, where $\beta = 2N_{e\beta 0}$ is the decay rate of expression divergence.

Author contributions: X.G. designed research; X.G. and Z.S. performed research; X.G. contributed new reagents/analytic tools; X.G. and Z.S. analyzed data; and X.G. and Z.S. wrote the paper.

The authors declare no conflict of interest.

Abbreviation: OU, Ornstein–Uhlenback.

‡To whom correspondence should be addressed at: 536 Science II Hall, Iowa State University, Ames, IA 50011. E-mail: xgu@iastate.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0610797104/DC1.

www.pnas.org/cgi/doi/10.1073/pnas.0610797104

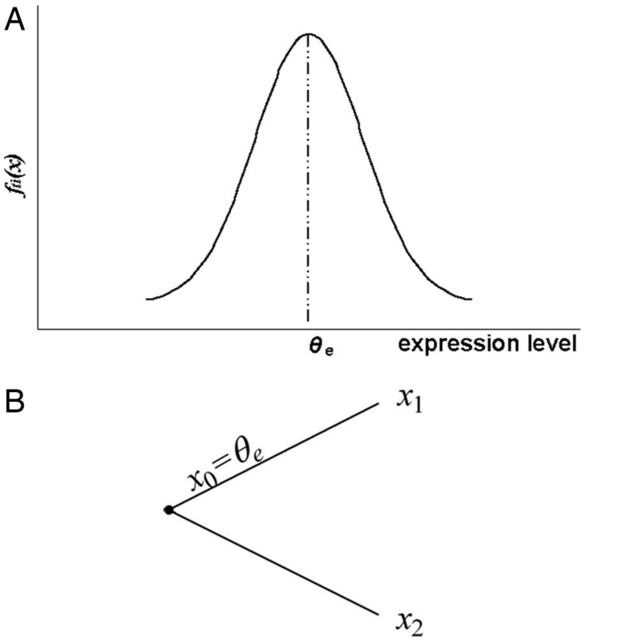

Fig. 1. Stabilizing model for tissue expression divergence. (*A*) Fitness function plotting against the expression level under the stabilizing selection; θ_e is the optimal expression level. (*B*) Scheme of interspecies expression divergence between two orthologous genes. Here, we assume that the ancestral expression level is at the optimal value ($x_0 = \theta_e$).

For two genomes, say, human and mouse, that have diverged t time units ago (Fig. 1), the expression distance can be derived similarly to Gu (12). Let x_1 and x_2 be the expression levels of two orthologous genes 1 and 2, respectively. Assuming the initial value is at the optimum ($x_0 = \theta_e$), from Eq. **2** we have $E[x_1 x_0] = E[x_2 x_0] = \theta_e$. If gene expression diverged along a lineage independently, we have $E[x_1 x_2 x_0] = E[x_1 x_0]E[x_2 x_0] = \theta_e{}^2$, and $E[x_1 x_2] = E[\theta_e{}^2]$, resulting in $\mathrm{Cov}(x_1,x_2) = Var(\theta_e)$. In the same manner, one can show $V(x_1) = V(x_2) = \epsilon^2(1 - e^{2\beta t})/2\beta + Var(\theta_e)$. Therefore, the expression distance for any gene pair g in tissue ti, $E_{ti,g} = E[(x_1 - x_2)^2]$, is given by

$$E_{ti,g} = \frac{\epsilon_g^2(1 - e^{-2\beta_g t})}{\beta_g} = \frac{(1 - e^{-2\beta_g t})}{W_{ti,g}}, \qquad [3]$$

where ϵ_g^2 is the mutational variance, β_g is the decay rate of expression divergence of gene pair g, and $W_{ti,g} = \beta_g/\epsilon_g^2$ is the strength of stabilizing selection on expression divergence. Thus, $E_{ti,g}$ is inversely related to $W_{ti,g}$. When $\rightarrow \infty$, $E_{ti,g} = 1/W_{ti,g}$.

Tissue-Dependent Evolutionary Rate of Protein Sequence. Gu (24) studied the evolutionary rate of a protein sequence, based on the principle that stabilizing selection on protein function generates sequence conservation. In the case of single protein function y (such as enzyme activity or DNA-binding affinity, also called molecular phenotype, the stabilizing selection on y follows a simple Gaussian form (Fig. 2)

$$f(y) = e^{-(y-\theta_g)2/2\sigma_g^2}. \qquad [4]$$

Thus, the coefficient of selection on y is given by $s(y) = 1 - f(y) \approx (y - \theta_g)^2/2\sigma_g^2$. On the other hand, random (nonsynonymous) mutations in the coding region affect the molecular phenotype y according to a distribution with the mean θ_g and variance $\sigma^2{}_m$ (Fig. 2). Consequently, the mean of selection of coefficient is given by $\bar{s} = E[(y - \theta_g)^2]/2\sigma_g^2 = \sigma_m^2/2\sigma_g^2$, and the selection intensity $S_g = 4N_e\bar{s} = 2N_e\sigma_m^2/\sigma_g^2$. In the general case of multiple (K) molecular phenotypes of protein function, Gu (24) have shown

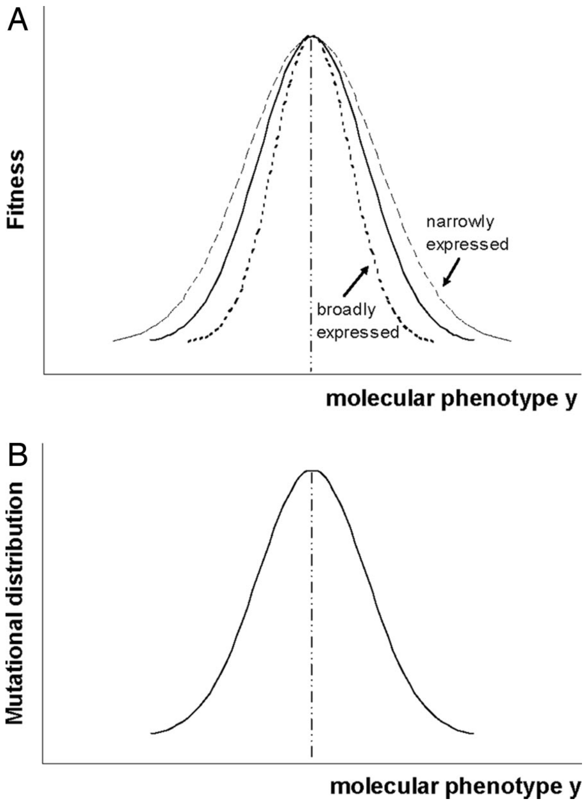

Fig. 2. The stabilizing model. (*A*) Fitness function plotting against the molecular phenotype (*y*) of protein function under the stabilizing selection. (*B*) Distribution of random mutations that affect the molecular phenotype *y*.

$$S_g = -2N_e \sum_{i=1}^{K} \frac{\sigma_{m,i}^2}{\sigma_{g,i}^2}, \qquad [5]$$

where the subscript i assigns σ_m^2 and σ_g^2 specific to the ith molecular phenotype.

Stabilizing selection of molecular phenotypes (measured by a set of $\sigma_{g,i}^2$) may be tissue-dependent, which can be modeled as $\sigma_{g,i}^2 = a_{g,i}^2/Z_g$, $i = 1,\ldots, K$. Whereas each $a_{g,i}^2$ is a tissue-independent constant, tissue factor Z_g measures the accumulated tissue effect on fitness; a larger Z_g means a greater tissue effect. For gene g expressed in L_g different tissues, we implement an additive-effect model $Z_g = \Sigma_{j=1}^{L_g} L_j$, in which L_j is the contribution from tissue j. The mean selection intensity in Eq. **5** can be rewritten in terms of tissue dependency

$$S_g = S_{g,0}Z_g, \qquad [6]$$

where $S_{g,0} = -2N_e\Sigma_{i=1}^{K}\sigma_{m,i}^2/a_{g,i}^2$ is the tissue-independent component. Hence, given the mutation rate v, the evolutionary rate of gene g is given by

$$\lambda_g = \upsilon\frac{S_g}{1 - e^{-S_g}} = \upsilon\frac{S_{g,0}Z_g}{1 - e^{-S_{g,0}Z_g}}. \qquad [7]$$

Eq. **7** links between-tissue-effects and evolutionary rate of protein sequence. Apparently, the evolutionary rate decreases when the accumulated tissue effect Z_g is strong and vice versa.

Tissue-Driven Hypothesis and Predictions. The tissue-driven hypothesis of genomic evolution postulates that the tissue factor plays an important role of functional constraint on the rate of genomic evolution, because genes influence phenotypic characters by expression in specific tissues. The phenotypic consequences of genetic variations in regulatory and coding sequences are both affected by the common microenvironment of tissues. Below, we discuss several predicted genomic correlations that can be tested by the genomic data.

Tissue Expression Distance (E_{ti}). To measure the expression difference of a tissue between two species, we define E_{ti} as the mean expression distance over N orthologous genes in tissue *ti*, that is, $E_{ti} = \sum_{g=1}^{N} E_{ti,g}/N$, where $E_{ti,g}$ is given by Eq. **3**. Under some moderate conditions, E_{ti} can be approximated by

$$E_{ti} \approx (1 - e^{-2\bar{\beta}t})/W_{ti} \qquad \textbf{[8]}$$

[see supporting information (SI) *Appendix*], where the mean tissue factor W_{ti} is the (harmonic) average of $W_{ti,g}$s, $\bar{\beta}$ is the mean decay-rate of expression divergence, and t is the time of speciation. Eq. **8** indicates that the tissue expression distance increases with time t and decreases with the mean tissue factor W_{ti}. When $\bar{\beta}$ is close to 0 (very weak stabilizing selection) or t is short (closely related species), Eq. **8** can be reduced to $E_{ti} \approx 2\epsilon^2 t$, i.e., the Brownian model (12, 13), where ε^2 is the mean mutational variance over genes. In the case of distantly related species when the expression divergence approaches the steady state, the time-dependent term in Eq. **8** vanishes, resulting in $E_{ti} \approx 1/W_{ti}$.

Tissue Expression and Sequence Distances: The $E_{ti} - D_{ti}$ Correlation. Let d_g be the evolutionary distance between an orthologous gene pair (g). For a set (N_{ti}) of genes that are expressed in tissue *ti*, the mean evolutionary distance is given by $D_{ti} = \sum_{g=1}^{N_{ti}} d_g/N_{ti}$. Because $d_g = 2\lambda_g t$, where λ_g is given by Eq. **7**, we have shown that the mean selection intensity of tissue (*ti*)-expressed protein sequences can be written as $\bar{S}_{ti} \approx S_0(Z_{ti} + \alpha)$ (see *SI Appendix*); Z_{ti} is the mean of accumulated tissue-(*ti*) factors over expressed genes, S_0 is the mean of tissue-independent components, and α is a constant. Thus, we have

$$D_{ti} \approx D_{\upsilon} \frac{S_{ti}}{1 - e^{-S_{ti}}}, \qquad \textbf{[9]}$$

where $D_{\upsilon} = 2\upsilon t$.

According to the tissue-driven hypothesis, two mean tissue factors W_{ti} and Z_{ti} should be positively correlated, because they represent the effects of common microenvironment of tissue *ti* on expression divergence and protein sequence conservation, respectively. This argument predicts a positive correlation between tissue expression distance (E_{ti}) and tissue sequence distance (D_{ti}). In the special case when $Z_{ti} = W_{ti}$ and $E_{ti} \approx 1/W_{ti}$ (steady-state expression divergence), we obtain the following form

$$\frac{D_{ti}}{D_{\upsilon}} \approx \frac{E_{ti}}{aE_{ti} + b}, \qquad \textbf{[10]}$$

where $a = 1 - S_{0\alpha}/2$ and $b = S_0/2$.

Interspecies and Interduplicate Tissue Expression Divergence: The $E_{ti} - T_{dup}$ Correlation. The tissue-driven hypothesis also predicts that tissue factors may affect the expression divergence between duplicate genes. Consider a pair of duplicate genes that have diverged τ evolutionary time units. Under the similar stabilizing selection model, one can obtain the expression distance between duplicated genes, $T_{dup,ti,g}$, which is virtually the same as Eq. **3**. To be clear, we use $Q_{ti,g}$ for the tissue factor of expression divergence between duplicate pair g. For a set (N_{dup}) of duplicate genes, let $T_{dup} = \sum_{g=1}^{N_{dup}} T_{dup,ti,g}/N_{dup}$ be the tissue (*ti*) duplicate distance. Similar to Eqs. **5** and **6**, we have

$$T_{dup} \approx (1 - e^{-2\bar{\gamma}\bar{\tau}})/Q_{ti}, \qquad \textbf{[11]}$$

where Q_{ti} is the mean tissue factor for the interduplicate expression divergence in tissue *ti*, $\bar{\gamma}$ is the mean decay rate of expression divergence, and $\bar{\tau}$ is the mean evolutionary time of the duplicate gene set. Hence, positively correlated W_{ti} and Q_{ti} under the tissue-driven hypothesis leads to a testable prediction of positive correlation between E_{ti} and T_{dup}. In particular, a linear $E_{ti} - T_{dup}$ relationship is expected when $W_{ti} = Q_{ti}$.

Tissue Broadness and Preference. One can rewrite the accumulated tissue effect on gene g in Eq. **6** as $Z_g = L_g \times \bar{Z}_g$, where L_g is the number of (L_g) of tissues in which gene g is expressed, and $\bar{Z}_g = \sum_{j=1}^{L_g} Z_j/L_g$ is the average tissue factor for gene g. In fact, $\bar{Z}_g$ measures the effect of tissue preference, or tissue types, on the expression divergence. In short, the accumulated tissue effect can be decomposed into two factors: tissue broadness (L_g) and tissue preference ($\bar{Z}_g$). The protein sequence becomes more conserved if the gene is expressed in more tissues or in tissues with more stringent constraints.

Although many studies have showed the effect of tissue broadness (9, 17, 25), the effect of tissue preference has not been well investigated. We address this issue by grouping genes with the same tissue broadness (L_g). When L_g is the same, the larger the $\bar{Z}_g$ value, the greater the selection intensity S_g and so the lower evolutionary rate λ_g. This prediction can be tested under the tissue-driven hypothesis that claims a positive correlation between $W_{ti,g}$ and $Z_{ti,g}$ (see below).

Results

In this section, we use the human and mouse genomic data to test these predicted genomic correlations derived from the tissue-driven hypothesis. We focused on 29 orthologous (adult) tissues in which the human and mouse microarrays are available; see *Materials and Methods*. For each tissue, we estimate the tissue expression distance (E_{ti}) between the human and mouse as well as the tissue protein distance (D_{ti}) and the tissue duplicate distance (T_{dup}) for expression divergence.

Tissue Expression Divergence Between Human and Mouse. Based on 8,936 human–mouse orthologs, we estimated the tissue expression distance E_{ti} for each of 29 tissues. Fig. 3 shows a substantial variation of E_{ti} among tissues. Indeed, there is a 2.4-fold

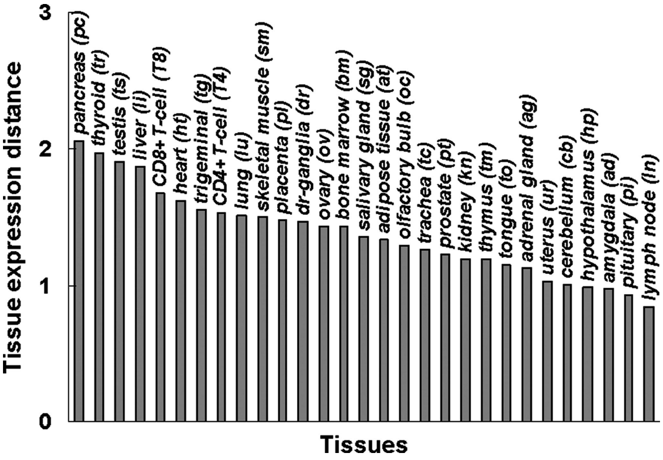

Fig. 3. Variation of human–mouse tissue expression distances (E_{ti}) among 29 tissues. Abbreviations for these tissues are shown in parentheses.

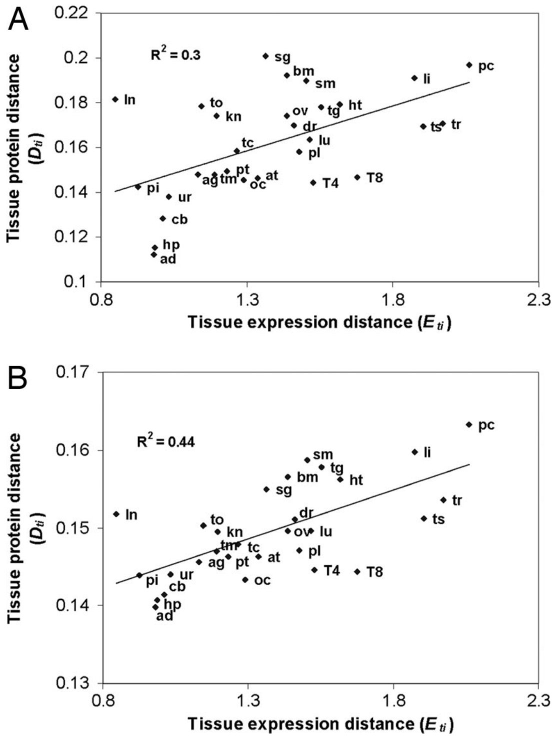

Fig. 4. Correlations between tissue expression distance (E_{ti}) and tissue protein distance (D_{ti}) for highly expressed proteins (*A*) and for normally expressed proteins (*B*). See Fig. 3 for the description of abbreviations of tissue names. In each case, the correlation is statistically highly significant ($P < 0.001$).

difference from the lowest $E_{\mathrm{ln}} = 0.85$ (lymph node, *ln*) to the highest $E_{\mathrm{pc}} = 0.206$ (pancreas, *pc*).

Previous studies (7, 8, 10) observed that brain may have more expression conservation than other tissues, but the small sample size (approximately five tissues) has raised some doubts. We addressed this issue because more (i.e., 29) tissues were examined. We found an overall expression conservation in some neurorelated tissues, e.g., pituitary (*pi*), amygdala (*ad*), hypothalamus (*hp*), and cerebellum (*cb*) (Fig. 3). In contrast, testis (*ts*) may have a rapid interspecies expression divergence. Although it remains open to question, one possibility is that the overall relaxed developmental constraint in the testis may facilitate the operation of sexual selection after speciation. Moreover, we noticed that some hormone-related tissues, such as pancreas, may have more developmental plasticity to allow rapid expression divergence, possibly through the interactions with environmental cues during evolution. In short, substantial variation of E_{ti} among tissues implies the role of tissue-specific factors in mammalian genomic evolution.

Correlation ($E_{\mathrm{ti}} - D_{\mathrm{ti}}$) Between Tissue Expression and Sequence Divergence. For each tissue *ti*, we calculated the tissue protein distance (D_{ti}) between the human and mouse. Similar to E_{ti}, the observed variation of D_{ti} among tissues may indicate the tissue's role in protein evolution. Moreover, the tissue-driven hypothesis expects covariation between E_{ti} and D_{ti}, because it postulates the same tissue-specific developmental constraint that may affect both tissue expression divergence and sequence divergence of expressed proteins. We indeed found a highly significant correlation between E_{ti} and D_{ti} based on 29 human–mouse tissues (Fig. 4). In the case of high expression (Fig. 4*A*), the (Pearson) coefficient of correlation is $R = 0.55$ ($P < 0.001$), whereas $R = 0.66$ ($P < 0.001$) in the case of normal expression (Fig. 4*B*). Use of the Spearman rank correlations results in very similar *P* values (<0.001). Hence, the significance of $E_{\mathrm{ti}} - D_{\mathrm{ti}}$ correlation provides statistical evidence to support the tissue-driven hypothesis. In addition to two cutoffs presented in Fig. 4 *A* and *B*, we have examined several other criteria for gene expression and found that the $E_{\mathrm{ti}} - D_{\mathrm{ti}}$ correlation is robust against the choice of cutoff (data not shown).

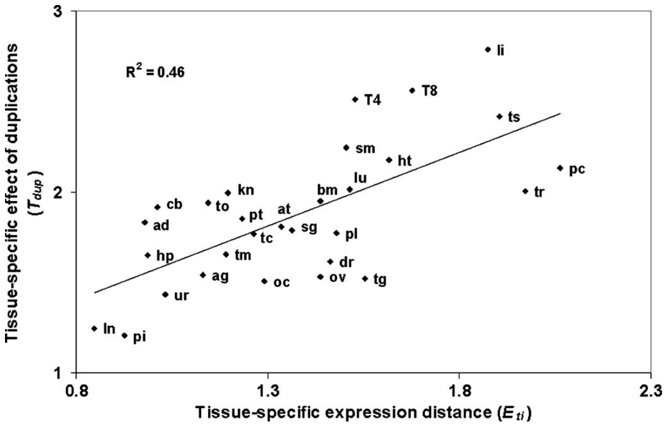

Fig. 5. The correlation between tissue expression distance (E_{ti}) and tissue duplicate distances (T_{dup}). Here, T_{dup} is the average of human and mouse duplicates. The correlation is statistically highly significant ($P < 0.001$).

Tissue Correlation ($E_{\mathrm{ti}} - T_{\mathrm{dup}}$) Between Interspecies and Duplicate Expression Divergence. A positive $E_{\mathrm{ti}} - T_{\mathrm{dup}}$ correlation implies that when a tissue allows more interspecies expression divergence, it should also tolerate more extensive expression divergence between duplicated genes. Based on 1,312 duplicate pairs that were duplicated before the human–mouse split, we estimated the duplicate tissue distance (T_{dup}) in each of tissue *ti*. Fig. 5 shows a highly significant correlation between tissue expression distance (E_{ti}) and T_{dup} ($P < 0.001$ for either Pearson or Spearman rank correlation). This result supports the tissue-driven hypothesis that duplicated genes tend to have more expression divergence in a tissue with relaxed developmental constraint and vice versa.

Evolutionary Rate of Protein Sequence Under Multiple Tissue Constraints. Let L_{g} be the number of tissues in which gene *g* is expressed, or the tissue broadness. For gene *g* that is expressed in L_{g} different tissues, we propose an index t_{g} that can be used to measure the effect of tissue preference approximately. We thus calculated the effect of tissue preference (t_{g}) for all 8,936 genes in both human and mouse. We further classified these genes into groups according to the number (L_{g}) of tissues in which they are expressed, i.e., $L_{\mathrm{g}} = 1, 2,\ldots 28$, excluding $L_{\mathrm{g}} = 29$ because t_{g} is identical in this case. Noticeably, for each group, a negative correlation between the protein distance (d_{g}) and t_{g} is observed (Fig. 6*A*). Twenty-five cases are statistically significant ($P < 0.05$), whereas cases of $L_{\mathrm{g}} = 11$, 14, and 23 are not ($0.05 < P < 0.1$) largely because of the small sample size. In particular, 15 cases show highly statistically significant ($P < 0.0001$). For instance, Fig. 6*B* shows the d_{g} vs. t_{g} correlation in the case of $L_{\mathrm{g}} = 5$. Here, we used AD = 200 as the cut-off for gene expression. For instance, we increased the cut-off up to AD = 800 to examine the so-called transcription leakage effect. At any rate, all these gave virtually the same results.

Given the same tissue broadness, the overwhelming negative $d_{\mathrm{g}} - t_{\mathrm{g}}$ correlation indicates that genes that are expressed in stronger constrained tissues (e.g., neurorelated) tend to evolve more slowly at the sequence level than those expressed in weaker constrained tissues (e.g., hormone-related), as predicted by the

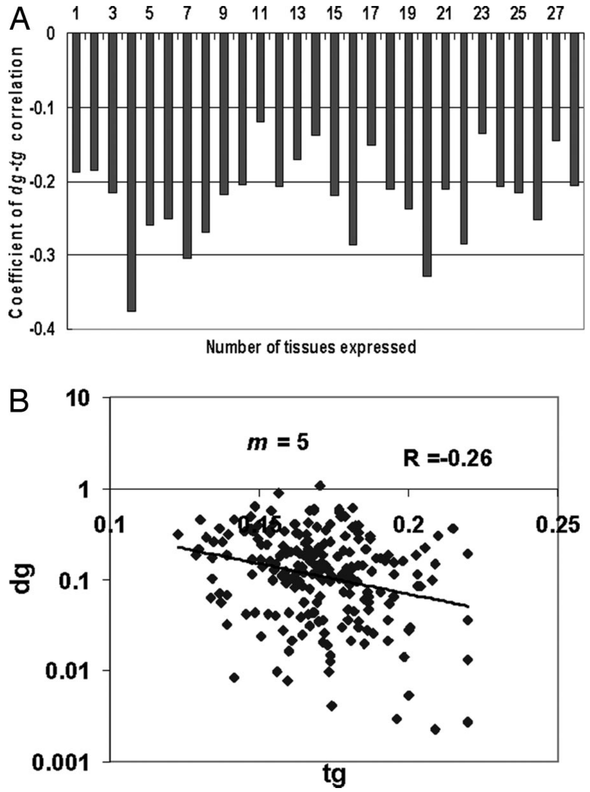

Fig. 6. Evolution under multiple tissue constraints. (*A*) Negative coefficients of $t_g - d_g$ correlations for gene groups with the same tissue broadness (L_g). (*B*) The $t_g - d_g$ plotting in the case of $L_g = 5$.

tissue-driven hypothesis. Apparently, if a protein is expressed in several different tissues, the evolution of protein sequence may be under multi-tissue-specific constraints. Hence, broadly expressed genes generally tend to evolve slowly at the sequence level. Indeed, we found a highly significant negative correlation between d_g and L_g, confirming previous studies (e.g., refs. 8 and 25) (data not shown).

Discussion

Under the stabilizing selection model, we have formulated the tissue-driven hypothesis and elaborated several predicted genomic correlations, taking advantage of multitissue human–mouse microarrays. In summary, we found highly significant correlations between tissue expression distance (E_{ti}) and tissue sequence distance (D_{ti}) and between E_{ti} and the duplicate tissue distance (T_{dup}), supporting the hypothesis that the evolution of expression pattern and protein sequence may be under the same constraint of tissue factors. Moreover, for genes with the same expression broadness, we found that genes expressed in more stringent tissues tend to evolve more slowly than those in more relaxed tissues. Our findings provide some insights on how the rate of genomic evolution can be shaped by the up-level physiological-developmental structure of the organism.

Functional Constraint vs. Positive Selection. A basic assumption of the tissue-driven hypothesis is that genome evolves largely under functional constraints maintained by stabilizing selections at levels from cell physiology to development. In some evolutionary lineages, episodic adaptive selection may happen either in expression pattern or in protein function (9, 21, 26–29). For instance, hundreds of genes (≈2% human genes) showed dramatic brain-specific expression shifts in the human lineage (3, 4, 26, 27). When the tissue-driven hypothesis is extended to include adaptive selection, we found the predictions for both $E_{ti} - D_{ti}$ and $E_{ti} - T_{dup}$ hold. We have examined the rapid-shift (S) model of expression divergence (12). In this case, one can show that the tissue expression distance in Eq. **8** can be modified as $E_{ti} = S_{hm} + (1 - e^{-2\bar{}\beta t})/W_{ti}$, and the duplicate tissue distance in Eq. **11** as $T_{dup} = S_{dup} + (1 - e^{-2\bar{}\beta t})/Q_{ti}$, where S_{hm} and S_{dup} are the rapid-shift components between human–mouse genes and between duplicate genes, respectively. Except for extreme cases, S_{hm} and S_{dup} apparently do not affect the predicted genomic correlations.

Effect of Expression Level on Protein Sequence Evolution. It has been claimed (9, 17, 18) that highly expressed genes tend to evolve slowly. We have examined this confound effect of tissue broadness and found that our main results are robust. For instance, high significance of $E_{ti} - D_{ti}$ correlation (Fig. 4) holds at various cutoffs, from normal to highly expressed genes. On the other hand, our model can be extended to take the effect of expression level into account, e.g., by assuming the tissue-factor Z_g is expression level-dependent.

Tissue Expression Pattern in Primates and Mammals. The $E_{ti} - D_{ti}$ correlation between the human and chimpanzee has been investigated by Khaitovich *et al.* (8), based on five tissues (brain, liver, heart, kidney, and testis). However, our reanalysis of the same data sets leads to nonsignificant result (Spearman rank test $P > 0.2$), as opposed to the original claim (8) (the Pearson correlation $P < 0.05$). It is known that the Pearson correlation could be too liberal in small sample size. Because the current study (29 human–mouse tissues) includes these five tissues, we did observe a roughly consistent ranking in E_{ti} or D_{ti}, i.e., the lowest values in the cerebellum/brain, whereas we found the highest values in the testis. Hence, one may speculate that the $E_{ti} - D_{ti}$ correlation holds in both primates and mammals, although more primate microarray data are needed.

Some Technical Issues. We have examined several technical issues that may affect our interpretations. First, our analysis is robust against the noise of microarrays, because the expression variation among biological replicates of microarrays is much smaller than the average expression difference between the human and mouse (16). Nevertheless, using the corrected expression distance (13), a conserved measure for interspecies expression divergence, we obtained virtually the same results (data not shown). Second, the exclusion of young duplicates (5) (after the human-mouse split) has almost no effect on our results. Third, we have used several alternative options to determine the status of expression level in a tissue. In all cases, highly significant genomic correlations are always observed.

Because of expression leakage or fluctuation, observed similar gene expression profiles do not necessarily mean a similar tissue function. The extent of these nonfunctional expressions is subject to the debate (30). It seems that the expression leakage may be more frequent in those tissues with relatively weak developmental constraints. Besides, evolution of expressions can be affected by many issues such as transregulatory elements (31) or the alternative splicing isoforms (32). Indeed, more questions are raised than we can solve in evolutionary genomics (30–33).

Materials and Methods

Genome Data Sets. Homology information of human and mouse genes was obtained from the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov/HomoloGene). After extracting the reciprocally unique hit pairs with IDs starting with the prefix "NM-," we identified a total of 17,462 high-quality human–mouse orthologous genes for further study. Meanwhile, human (HG-U133A and GNF1H) and mouse (GNF1M) Affymetrix microarray data (Affymetrix, Santa Clara, CA) were retrieved from http://

symatlas.gnf.org (16). We focused on the following 29 orthologous (adult) tissues in which the human and mouse microarrays are available: Adipose tissue (*at*), adrenal gland (*ag*), amygdala (*ad*), bone marrow (*bm*), cerebellum (*cb*), CD4$^+$ T cells (*T4*), CD8$^+$ T cells (*T8*), dorsal root ganglion (*dr*), heart (*ht*), hypothalamus (*hp*), kidney (*kn*), liver (*li*), lung (*lu*), lymph node (*ln*), olfactory bulb (*oc*), ovary (*ov*), pancreas (*pc*), pituitary (*pi*), placenta (*pl*), prostate (*pt*), salivary gland (*sg*), skeletal muscle (*sm*), testis (*ts*), thymus (*tm*), thyroid (*tr*), tongue (*to*), trachea (*tc*), trigeminal (*tg*), and uterus (*ur*). As suggested by the authors (16), we mainly used the normalized (log2-based) ratio value (AffyRatio) of the medium expression value among biological replicates. Using the annotation tables at http://symatlas.gnf.org, we matched the human–mouse orthologous genes to the human and mouse Affymetrix tags, respectively. The final data set included 8,936 human–mouse orthologous genes. Note that ≈20% of cases that had multiple tags in the microarray were targeted against the single gene. We solved this problem by assigning the averaged or the highest expression value for each of these genes (16). Nevertheless, these two treatments provided virtually the same results.

Estimation of Tissue Expression Distance (E_{ti}). Consider a set (N) of orthologous genes between species 1 (human) and species 2 (mouse). Let $x_{g1,ti}$ and $x_{g2,ti}$ be the (log2-transformed) expression levels of the gth orthologous genes in tissue *ti*, respectively. Under the OU model, one can easily show that the tissue (*ti*) expression distance defined in Eq. **8** can be estimated as follows

$$\hat{E}_{ti} = \sum_{g=1}^{N} (x_{g1,ti} - x_{g2,ti})^2/N. \qquad \textbf{[12]}$$

Estimation of Tissue Protein Distance (D_{ti}). We calculated D_{ti} as the mean of evolutionary distances of proteins that are expressed in tissue *ti*. For each gene, the evolutionary distance was estimated by the Poisson correction; other methods gave virtually the same results (data not shown). For each tissue *ti*, we inferred the status of gene expression as follows: (*i*) High expression: the AffyRatio of the gene is above the medium expression among 79 human tissues (16). (*ii*) Normal expression: calculate the percentages of AD counts (adjusted by the background AD = 200) of the gene in all 29 tissues and then, in a descending order, select the expressed tissues of the gene until the accumulated AD percentage up to 97.5%. This approach may avoid some spurious high AD counts.

Estimation of Tissue Duplicate Distance (T_{dup}) for Expression Divergence. Consider a set (N_{dup}) of duplicate gene pairs. For the jth duplicate pair, the expression levels (AffyRatio) of two duplicate genes in a given tissue (*ti*) are denoted by x_j and y_j, respectively. Then, similar to the calculation of E_{ti} in Eq. **12**, we estimate T_{dup} by the formula

$$\hat{T}_{dup} = \sum_{j=1}^{N_{dup}} (x_j - y_j)^2/N_{dup}. \qquad \textbf{[13]}$$

A large T_{dup} value reflects the plasticity of tissue-specific developmental constraint that allows more expression divergence between duplicate genes.

Estimation of Tissue Broadness and Preference. The number (L_g) of tissues in which gene g is expressed, or the tissue broadness, can be inferred as described above. For gene g that is expressed in L_g different tissues, let E_j ($j = 1,\ldots, L_g$) be the jth tissue expression distance between the human and mouse. Because a large E_j means less tissue constraint on expression divergence, we propose an index that can be used to measure the effect of tissue preference as follows

$$t_g = \sum_{j=1}^{L_g} E_j^{-1}/L_g, \qquad \textbf{[14]}$$

where tissue expression distance E_j is estimated by Eq. **12**. In particular, when the expression divergence is close to the steady state, we have $E_j \approx 1/W_j$ so that t_g is an estimate of the mean tissue factor $\sum_{j=1}^{L_g} W_j/L_g$, which is a proxy for the tissue preference $\bar{Z}_g = \sum_{j=1}^{Lg} Z_j/L_g$ under the tissue-driven hypothesis that predicts $W_j \approx Z_j$, creating a negative correlation between t_g and the evolutionary distance of protein sequence (d_g).

We thank Dongping Xu for computer assistance. This work was supported by a National Institutes of Health grant and the National Science Foundation of China Overseas Outstanding Young Investigator Award (to X.G.).

1. Kimura M (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ Press, Cambridge, UK).
2. Wilson AC, Carlson SS, White TJ (1977) *Annu Rev Biochem* 46:573–639.
3. Enard W, Khaitovich P, Klose J, Zollner S, Heissig F, Giavalisco P, Nieselt-Struwe K, Muchmore E, Varki A, Ravid R, *et al.* (2002) *Science* 296:340–343.
4. Gu J, Gu X (2003) *Trends Genet* 19:63–65.
5. Huminiecki L, Wolfe KH (2004) *Genome Res* 14:1870–1879.
6. Khaitovich P, Weiss G, Lachmann M, Hellmann I, Enard W, Muetzel B, Wirkner U, Ansorge W, Paabo S (2004) *PLoS Biol* 2:E132.
7. Yanai I, Graur D, Ophir R (2004) *Omics* 8:15–24.
8. Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, Weiss G, Lachmann M, Paabo S (2005) *Science* 309:1850–1854.
9. Liao BY, Zhang J (2006) *Mol Biol Evol* 23:530–540.
10. Gu Z, Nicolae D, Lu HH, Li WH (2002) *Trends Genet* 18:609–613.
11. Makova KD, Li WH (2003) *Genome Res* 13:1638–1645.
12. Gu X (2004) *Genetics* 167:531–542.
13. Gu X, Zhang Z, Huang W (2005) *Proc Natl Acad Sci USA* 102:707–712.
14. Li WH, Yang J, Gu X (2005) *Trends Genet* 21:602–607.
15. Gilad Y, Oshlack A, Rifkin SA (2006) *Trends Genet* 22:456–461.
16. Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, *et al.* (2002) *Proc Natl Acad Sci USA* 99:4465–4470.
17. Duret L, Mouchiroud D (2000) *Mol Biol Evol* 17:68–74.
18. Yang J, Su AI, Li WH (2005) *Mol Biol Evol* 22:2113–2118.
19. Rifkin SA, Houle D, Kim J, White KP (2005) *Nature* 438:220–223.
20. Denver DR, Morris K, Streelman JT, Kim SK, Lynch M, Thomas WK (2005) *Nat Genet* 37:544–548.
21. Jordan IK, Marino-Ramirez L, Koonin EV (2005) *Gene* 345:119–126.
22. Lande R (1979) *Evolution (Lawrence, Kans)* 33:234–251.
23. Hansen TF, Martins EP (1996) *Evolution (Lawrence, Kans)* 50:1404–1417.
24. Gu X (November 1, 2006) *Genetica*, 10.1007/s/0709-006-0022-5.
25. Zhang L, Li WH (2004) *Mol Biol Evol* 21:236–239.
26. King MC, Wilson AC (1975) *Science* 188:107–116.
27. Caceres M, Lachuer J, Zapala MA, Redmond JC, Kudo L, Geschwind DH, Lockhart DJ, Preuss TM, Barlow C (2003) *Proc Natl Acad Sci USA* 100:13030–13035.
28. Piganeau G, Eyre-Walker A (2003) *Proc Natl Acad Sci USA* 100:10335–10340.
29. Sella G, Hirsh AE (2005) *Proc Natl Acad Sci USA* 102:9541–9546.
30. Yanai I, Korbel JO, Boue S, McWeeney SK, Bork P, Lercher MJ (2006) *Trends Genet* 22:132–138.
31. Zhang Z, Gu J, Gu X (2004) *Trends Genet* 20:403–407.
32. Su Z, Wang J, Yu J, Huang X, Gu X (2006) *Genome Res* 16:182–189.
33. Eisen JA, Fraser CM (2003) *Science* 300:1706–1707.