Annals of Biomedical Engineering, Vol. 35, No. 6, June 2007 (© 2007) pp. 1088–1094
DOI: 10.1007/s10439-007-9269-y

Tissue-driven Hypothesis with Gene Ontology (GO) Analysis

ZHIXI SU,[1] YONG HUANG,[2] and XUN GU [1,2]

[1]Institutes of Biomedical Sciences, School of Life Sciences, Center for Evolutionary Biology, Fudan University, Shanghai, China and [2]Department of Genetics, Development, and Cell Biology, Center for Bioinformatics and Biological Statistics, Iowa State University, 536 Science II hall, Ames, IA 50011, USA

(*Received 28 September 2006; accepted 25 January 2007; published online 20 March 2007*)

Abstract—Most of the genes are under selective pressure to maintain their expression levels in the tissues. In a recent study, we have proposed a "tissue-driven" hypothesis stating that the stabilizing constraints on gene expression levels can be partitioned among tissues; tissues differ in their tolerance to gene expression variances; and the constraints on expression divergence is correlated with the constraints on sequence divergence. Here we further tested the "tissue-driven" hypothesis by sub-grouping genes into Gene Ontology (GO) categories. We examined the distribution of tissue expression distance of genes in the major GO categories in the tissues. We also examined the correlation between tissue expression distances and tissue sequence distances or tissue duplicate distances in the major GO categories. Our results have shown that the tissues-specific stabilizing constraints are generally not dominated by particular GO categories. It is also shown that sub-grouping genes into GO categories increased the sensitivity for detecting potential positive factors in expression divergence in the tissues.

Keywords—Expression divergence, Tissue expression distance, Tissue sequence distance, Tissue duplicate distance.

INTRODUCTION

Mutations can influence the phenotype of genes at two levels, the coding sequence level and the gene expression level. Actually, it has long been postulated that between the species with highly similar gene sequences, such as human and chimpanzee, the differences in gene expression level may account for most of the phenotypic differences.[15] In recent years, the advance in microarray technology has made genome-wide expression data of many species available, which greatly facilitates the analysis of gene expression divergence in evolution.

Many studies have investigated the expression divergence between species, which are summarized in several recent reviews.[4,10,23] Briefly, large number of quantitative changes in gene expression is found among primates in the initial study,[2] but follow-up studies had somewhat conflicting results. Some studies support a neutral model of expression divergence,[11,12,14,25] while other studies[5,9,17] showed the general existence of stabilizing constraints. Some studies have also found positive selection in primate or mammal expression divergence.[5,13] There are also many studies on expression divergence in non-mammal species,[3,16,18–20,24] most support a stabilizing selection model but do not exclude the existence of neutral or positive factors.

Although there are still debates on the details of the evolution models, we recognized that most of the current results are consistent with the stabilizing selection model (allowing the neutral factor). Meanwhile, the stabilizing selection model is also a valid null hypothesis for testing positive selection. In a recent study by Gu and Su,[7] we proposed a statistical model of gene expression divergence driven by stabilizing selection. It is also proposed in that study a "tissue-driven" hypothesis stating that the selective pressure on the evolution of sequence and expression are correlated and can be partitioned among tissues, and tissues differ in their tolerance to gene expression variances.

We also recognized that individual genes may be under different selective pressures, owing to their particular properties. The Gene Ontology (GO) categorization[8] is arguably the most comprehensive summarization of functional and positional properties of the genes. In this report, we further tested the "tissue-driven" hypothesis by sub-grouping the genes into the GO categories. We examine the distribution of the tissue expression distance of the major GO categories in the tissues. We also examined the correlation between tissue expression distance and tissue sequence distance or tissue duplicate distance in the major GO categories. We found that the tissue-specific constraints

Zhixi Su and Yong Huang contributed equally to this work.

Address correspondence to Xun Gu, Department of Genetics, Development, and Cell Biology, Center for Bioinformatics and Biological Statistics, Iowa State University, 536 Science II hall, Ames, IA 50011, USA. Electronic mail: xgu@iastate.edu

on expression divergence are not dominated by any particular GO categories. On the other hand, subgrouping genes into GO categories has improved the resolution and sensitivity of expression divergence analysis.

MATERIALS AND METHODS

Datasets

Homology information of human and mouse genes was obtained from NCBI Homologene (http://www.ncbi.nlm.nih.gov/HomoloGene). Only the reviewed RefSeq genes are used and all the orthologous pairs are reciprocally unique. A total of 17,462 high-quality human–mouse orthologous pairs were identified. The expression information were retrieved from the GNF SymAtlas (http://symatlas.gnf.org).[21,22] The data were from the human HG-U133A and GNF1H chips, and the mouse GNF1M chip. Our study was focused on 29 orthologous (adult) tissues or cell types (referred to as ‚tissues' for simplicity) shared by the mouse and human expression datasets, which are listed as follows: adipose tissue (*at*), adrenal gland (*ag*), amygdala (*ad*), bone marrow (*bm*), cerebellum (*cb*), $CD4^+$ Tcells (*T4*), $CD8^+$ Tcells (*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*); the abbreviations in the parentheses were used in the context and figures. As suggested by the original authors,[22] we used the normalized ($\log_2$-based) ratio value (AffyRatio) of the median expression value among biological replicates. Using the annotation tables available at http://symatlas.gnf.org, we mapped the human–mouse orthologous genes onto the human and mouse Affymatrix tags. The final dataset included 8936 human–mouse ortholog pairs with expression information.

Tissue Expression Distance (E_{ti})

Consider a set of n orthologous gene pairs between species 1 (human) and species 2 (mouse). Let $x_{1,g,ti}$ and $x_{2,g,ti}$ denote the (log2-transformed) expression levels of the genes in the g-th pair in tissue *ti*. The mean tissue expression distance $\bar{E}_{ti}$ is calculated as

$$\bar{E}_{ti} = \sum_{g=1}^{n} (x_{1,g,ti} - x_{2,g,ti})^2 / n \qquad (1)$$

For genes in the GO category of *cat*, the mean expression distance in tissue *ti* is

$$\bar{E}_{ti,cat} = \sum_{g=1}^{n_{cat}} (x_{1,g,ti} - x_{2,g,ti})^2 / n_{cat}, \qquad (2)$$

where n_{cat} is the number of orthologous gene pairs in the *cat* category.

Tissue Sequence Distance (D_{ti})

For the human genes that are expressed in tissue *ti*, we calculated the mean tissue sequence distances D_{ti} as the mean evolutionary distance between these genes and their mouse orthologs.

$$\bar{D}_{ti} = \sum_{g=1}^{n} -\ln(I_g/100)/n, \qquad (3)$$

where n is the number of human genes expressed in tissue *ti*, and I_g is the sequence identity percentage (ranging from 0–100) of orthologous gene pair g between human and mouse. For the genes in GO category *cat*, the mean tissue sequence distance is

$$\bar{D}_{ti,cat} = \sum_{g=1}^{n_{cat}} -\ln(I_g/100)/n_{cat}, \qquad (4)$$

where n_{cat} is the number of the human genes expressed in tissue *ti* and in GO category *cat*.

For each orthologous pair, the evolutionary distance was estimated with the Poisson-correction. For a human gene, the tissues that has ‘‘High expression'' is defined as the tissues where the expression level of the gene is above its median expression level among all 79 human tissues in the dataset; the tissues that the gene has ‘‘Normal expression'' is defined as the minimum set of tissues to have an accumulative expression reading exceeding 97.5% of the sum of expression readings in all the 79 human tissues. Our definition of ‘‘Normal expression'', in effect, excludes the tissues with trace expression (though still above the background) from being counted as expressing tissues. By these definitions, the expression level of a gene in a tissue is compared to its expression in other tissues, rather than to other genes in the same tissue.

Tissue Duplicate Distance (T_{dup})

Duplicated genes were identified based on the methods in Gu *et al.*[6] 2698 human duplicate pairs were identified. To analyze the correlation between tissue expression distance (E_{ti}) and tissue duplication distance (T_{dup}), we further selected the duplicated pairs (1312 pairs) that have duplicated before the human–mouse split. A duplicated human gene pair is deemed

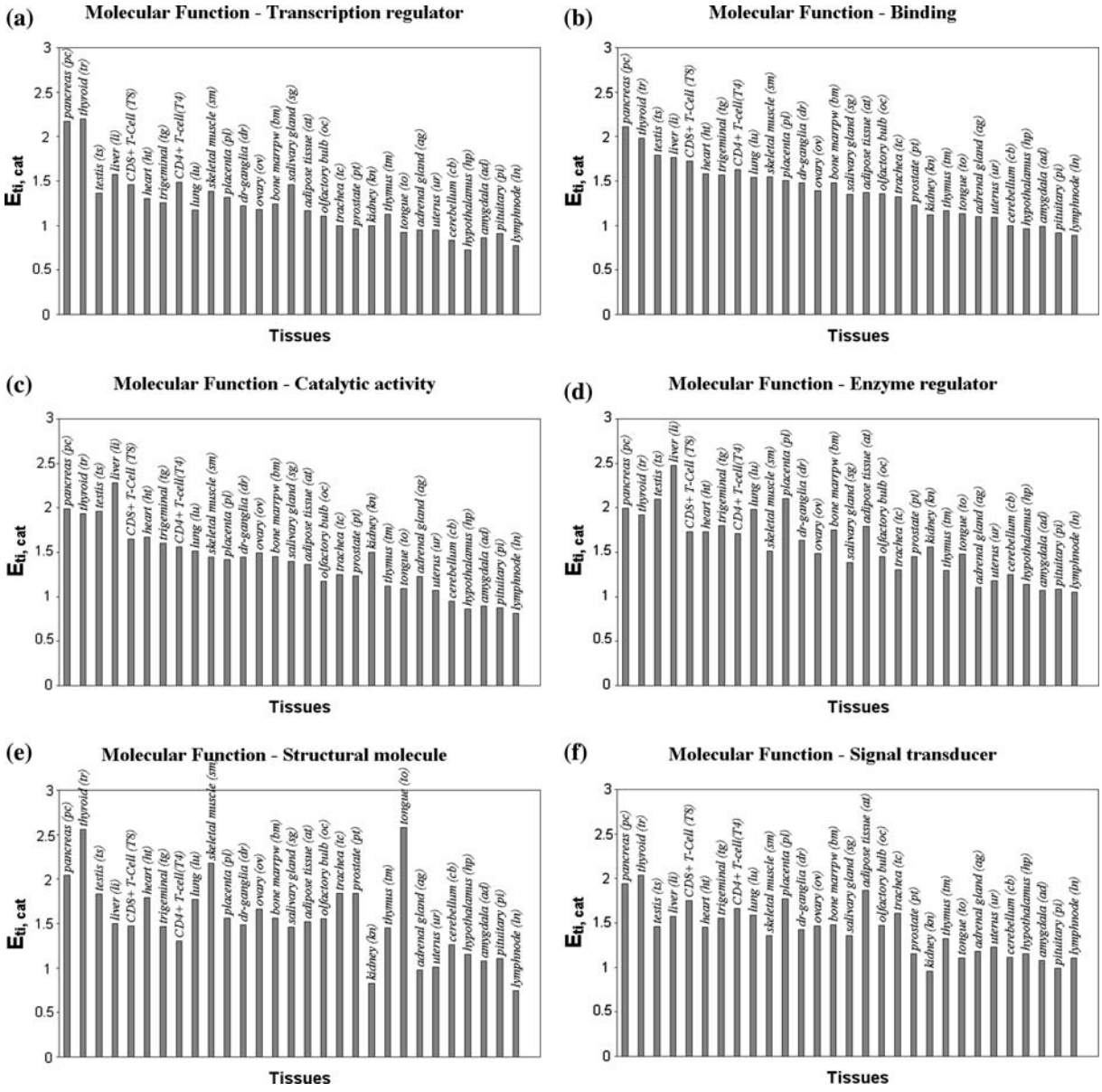

FIGURE 1. The distribution of the mean tissue expression distances ($E_{ti,cat}$) in different GO "molecular function" categories. (a) Transcription regulator; (b) Binding; (c) Catalytic activity; (d) Enzyme regulator; (e) Structural molecule; and (e) Signal transducer.

to have duplicated before the human–mouse split, if each gene in the pair has a reciprocally unique orthologous gene in mouse and the distance between the human duplicated genes is longer than the correspondent human–mouse orthologs. We excluded the duplicated pairs that duplicated after the human–mouse split in the study, because the tissue expression distance (E_{ti}) is calculated from human–mouse orthologous genes. Including the recently duplicated gene pairs would bring bias into the analysis.

Consider a set of m duplicated gene pairs. For the j-th pair, the expression levels of two duplicate genes in a given tissue (ti) are denoted as x_j and y_j, respectively. Thus, the mean tissue duplicate distance in a given tissue can be calculated as

$$\bar{T}_{dup} = \sum_{j=1}^{m} (x_j - y_j)^2 / m \tag{5}$$

For the genes in GO category cat, the mean tissue duplicate distance is

$$\bar{T}_{dup,cat} = \sum_{j=1}^{m_{cat}} (x_j - y_j)^2 / m_{cat}, \tag{6}$$

where m_{cat} is the number of duplicate gene pairs in GO category cat.

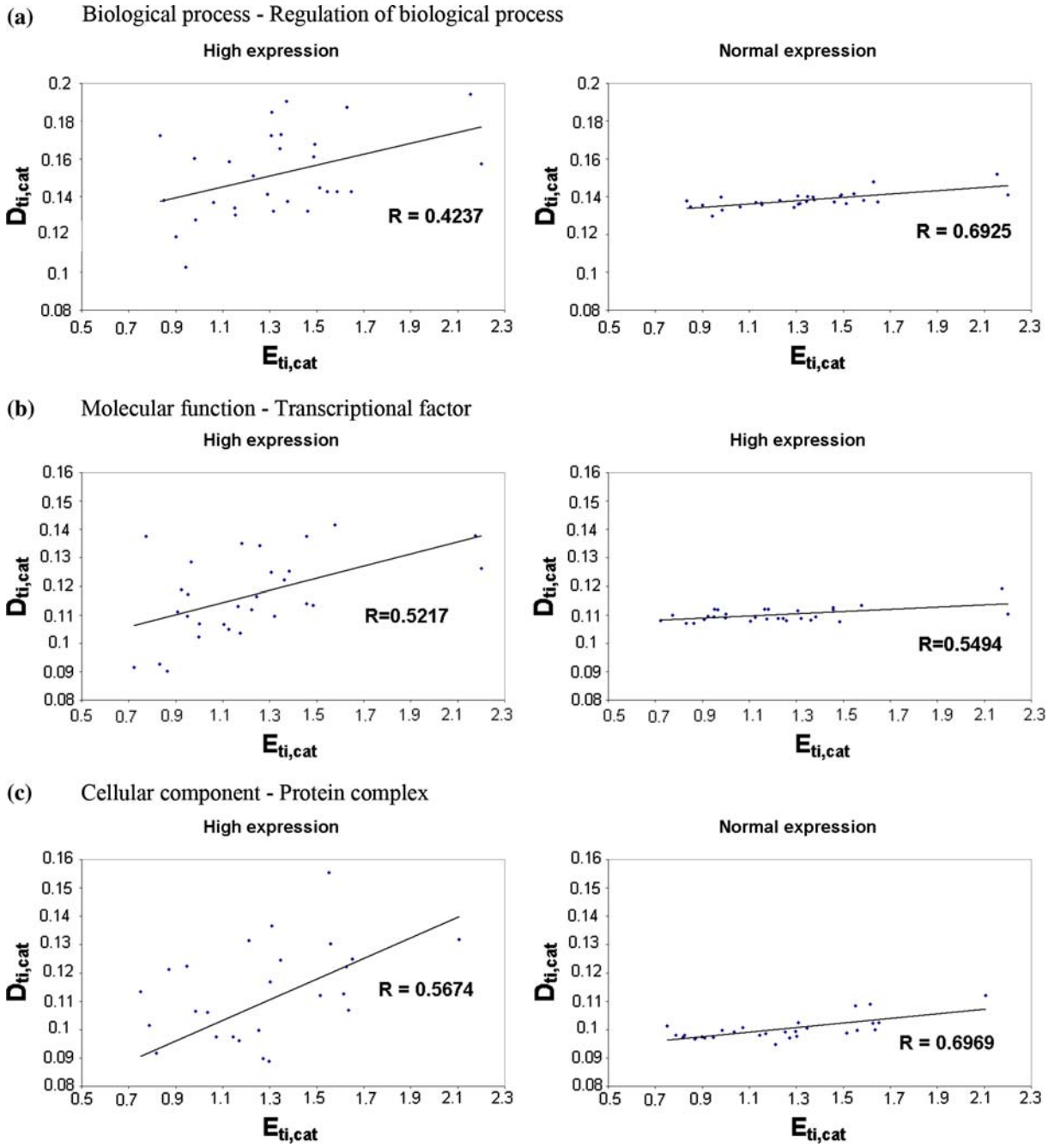

FIGURE 2. The correlation between $E_{ti,cat}$ and $D_{ti,cat}$ for both genes of "High expression" and "Normal expression" in several selected GO categories.

Gene Ontology (GO) Association of Genes

Gene ontology was downloaded from the GO consortium (http://www.geneontology.org/).[1] Gene Ontology association was based on NCBI LocusLink (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene). We used the major categories defined by the "generic GO slim",[8] which offered a high level view of the ontologies. Customized PERL scripts were written to map the genes onto the categories, with considerations of the DAG (directed acyclic graph) nature of the GO terms.

RESULTS

Tissue Expression Distance between Human and Mouse in Different GO Categories

We calculated the mean tissue expression distance $\bar{E}_{ti,cat}$ in major GO categories between human and

mouse. In Fig. 1 we displayed the histogram of the $\bar{E}_{ti,cat}$ values of several major GO "molecular function" categories in the tissues (see Supplement for other GO categories). The tissues are ordered in the same order as is in Fig. 3 of Gu and Su[7] (the general distribution), where the mean tissue expression distances $\bar{E}_{ti}$ are in descending order among the tissues. Comparison of the histograms of $\bar{E}_{ti,cat}$ and $\bar{E}_{ti}$ displays the deviation of individual $\bar{E}_{ti,cat}$ distribution from the general distribution. As is shown in Fig. 1, the distribution in individual GO categories is consistent with the general distribution in general. This is particular obvious in the categories of the more general functions such as binding and transcription regulator. However, there are also many deviations in some GO categories showing that the constraint on expression divergence is diversified among GO categories. For example, in the GO categories of "Catalytic activity" and "Enzyme regulator activity", the expression divergence is significantly increased in liver between human and mouse. Another interesting deviation is in the "Structural molecule activity" category, the expression divergence is most profound in the tissue of tongue between human and mouse. These deviations are good indicators of turbulences in the evolution of the expression levels of genes in these categories, although more refined GO annotation and molecular biology experiments would be needed for further analysis. These results have shown that adding the GO information has greatly improved the resolution and sensitivity of expression divergence analysis.

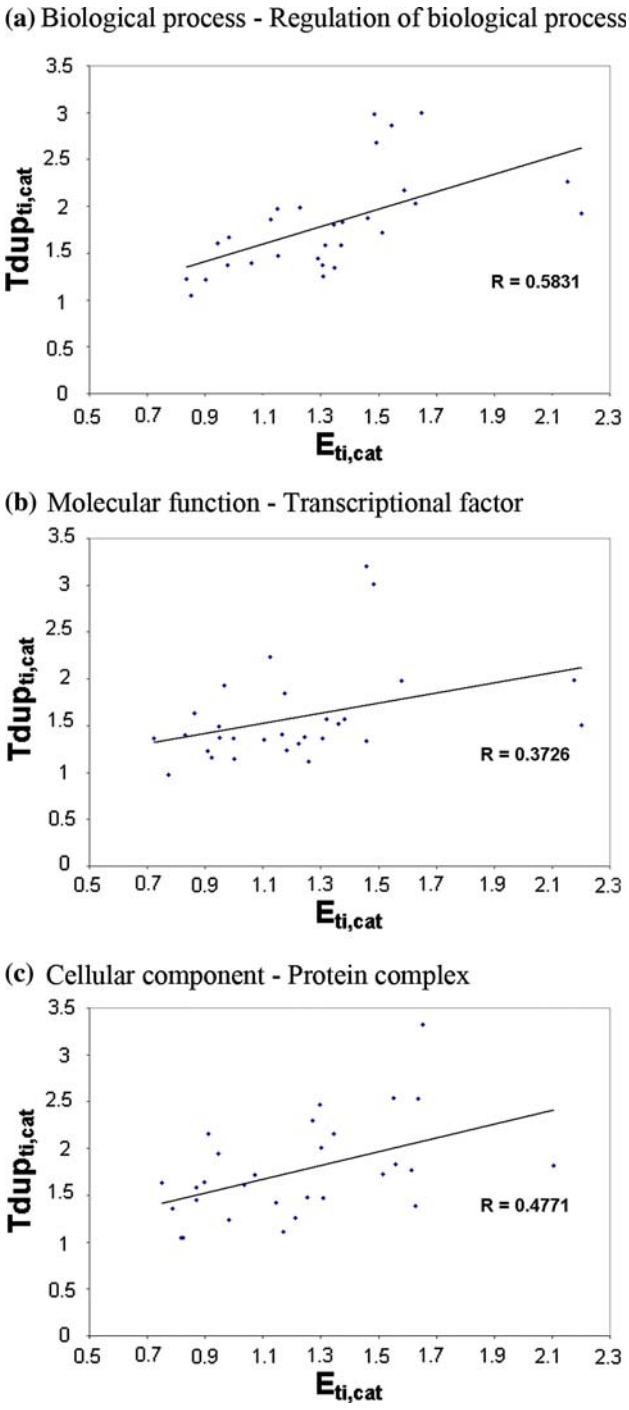

FIGURE 3. The correlation between $E_{ti,cat}$ and $T_{dup,cat}$ in several selected GO categories.

Correlation between Tissue Expression Distance and Tissues Sequence Distance in Major GO Categories

In each tissue *ti*, we calculated the mean tissue sequence distance for the major GO categories ($D_{ti,cat}$), for "High expression" and "Normal expression" genes respectively. The correlation between $\bar{E}_{ti,cat}$ and $D_{ti,cat}$ are calculated for the major GO categories in the 29 tissues in the dataset. The result is summarized in Table 1. It can be seen from the table that, except for the categories of "biological process unknown", "molecular function unknown", "extracellular matrix" and "extracellular region", the correlation for $\bar{E}_{ti,cat}$ and $D_{ti,cat}$ are generally significant in the GO categories. Meanwhile, except for the "structural molecule activity" category, the significance of the correlation does not differ much between the "High expression" genes and "Normal expression" genes. Although, the $D_{ti,cat}$ usually have larger variance in the "High expression" genes than the "Normal expression" genes. In Fig. 2, the plots of $D_{ti,cat}$ *vs.* $\bar{E}_{ti,cat}$ in the "Regulation of biological process", "Transcriptional factor" and "Protein complex" GO categories were shown as examples.

Although a strong correlation between the expression divergence and the sequence divergence is not a direct indicator of either the negative or the positive selection model, it however shows that there is common evolutionary constraint on expression divergence and sequence divergence. In our previous study,[7] we

TABLE 1. Summary of gene ontology analysis.

	GO Category	Number of orthologs	Number of duplicate pairs	r(Eti-Tdup)	r(Dti-Eti):High expression	r(Dti-Eti):Normal expression
Biological Process	Biological_process unknown	157	17	0.25	0.25	0.32
	Cellular process	5621	809	0.64**	0.54**	0.70**
	Development	1064	169	0.40*	0.33	0.35
	Physiological process	5453	781	0.68**	0.55**	0.69**
	Regulation of biological process	1462	208	0.58**	0.42*	0.69**
Cellular Component	Cell	5175	726	0.58**	0.52**	0.70**
	Cellular_component unknown	196	22	0.42*	0.40*	0.44*
	Extracellular matrix	203	38	0.31	0.02	0.08
	Extracellular region	628	97	0.70**	0.17	0.23
	Organelle	2875	345	0.60**	0.64**	0.80**
	Protein complex	813	97	0.48**	0.57**	0.70**
Molecular Function	Transporter activity	832	153	0.42*	0.38*	0.36
	Transcription regulator activity	624	89	0.37*	0.52**	0.55**
	Binding	4215	591	0.62**	0.51**	0.68**
	Catalytic activity	2481	364	0.82**	0.63**	0.69**
	Enzyme regulator activity	297	40	0.63**	0.53**	0.37*
	Molecular_function unknown	177	10	0.48**	0.16	0.41*
	Signal transducer activity	1278	210	0.42	0.36	0.42*
	Structural molecule activity	339	61	0.67**	0.45*	0.18

Note: *, ** indicate the significance levels of 0.05 and 0.01 respectively

have shown this constraint exists on the tissue level. Here, we further show that this correlation does not break when the genes expressed in a tissue are sub-grouped into GO categories. This shows that the stabilizing effect of the tissues is not dominated by genes in particular GO categories.

Correlation between Tissue Expression Distance and Tissues Duplicate Distance in View of GO Categories

In each tissue *ti*, the mean tissue duplicated distances are calculated for the major GO categories ($\bar{T}_{dup,cat}$), and the correlation between $\bar{E}_{ti,cat}$and $\bar{T}_{dup,cat}$are calculated. The result is summarized in Table 1. In most of the GO categories, $\bar{E}_{ti,cat}$ shows a significant correlation with $\bar{T}_{dup,cat}$. In Fig. 3, the plots of $\bar{T}_{dup,cat}$ vs. $\bar{E}_{ti,cat}$ in the ''Regulation of biological process'', ''Transcriptional factor'' and ''Protein complex'' GO categories were shown as examples. The strong correlation shows that in the tissues where the stabilizing constraint for expression divergence between species is weaker, its tolerance to expression divergence between duplicated genes is also larger; and *vice versa*. In the previous study,[7] we have shown this on a tissue level. Here we show that sub-grouping the genes by their GO categories does not break this correlation. This also shows that this effect is not dominated by genes in particular GO categories.

DISCUSSION

In this study, we sub-grouped genes into the major GO categories and further tested the ''tissue-driven'' hypothesis. Our result has shown that, in most of the situations, the tissue-specific constraints are also reflected in the individual GO categories. This is particular obvious in the neural tissues and the hormone-related tissues. The mean tissue expression distances in the neural tissues, such as pituitary (pi), amygdala (ad), hypothalamus (hp), and cerebellum (cb), are always among the lowest in most of the GO categories examined. Meanwhile, the mean tissue expression distance in the hormone-related tissues, such as pancreas (pc) and thyroid (tr), are always among the highest, also in most of the GO categories examined. This has shown that the tissue-specific stabilizing constraint is not dominated by genes in particular GO categories. The tissue-specific stabilizing constraint is more likely to be governed on more general levels, such as the interaction between the tissues (adult or embryonic) or the physiological structure of the tissues.

On the other hand, in some GO categories, the mean tissue expression distance showed interesting deviation from the general distribution between human and mouse, for example, the increased mean tissue expression distances in the ''Structural molecule'' category in the tissue of tongue, and in the ''catalytic activity'' category in the tissue of liver. This shows that in particular tissue and in particular GO category, potential indications of positive selection do exist between species. However, we have not found such indicative deviations in neural tissues in the GO categories. This has shown that the constraint on gene expression variance in the neural tissue is particularly strong. This is contradictory to the vast phenotypical differences in these tissues between human and mouse.

However, since all the expression data are from adult tissues, it is possible that divergence in expression among these tissues may be profound in other developmental stages, such as embryogenesis. The tissue-specific stabilizing constraints we shown here and in the previous report[7] are based on the adult tissue, although the whole set of methods can be readily applied to tissues in other developmental stages once the data is available.

ACKNOWLEDGMENT

This work has been supported by NIH grants to X. G.

REFERENCES

[1]Ashburner, M., C. A. Ball, and J. A. Blake et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*. 25:25–29, 2000.

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

[3]Fay, J. C., H. L. McCullough, P. D. Sniegowski, and M. B. Eisen. Population genetic variation in gene expression is associated with phenotypic variation in Saccharomyces cerevisiae. *Genome Biol*. 5:R26.

[4]Gilad, Y., A. Oshlack, and S. A. Rifkin. Natural selection on gene expression. *Trends. Genet*. 22:456–461, 2006.

[5]Gilad, Y., A. Oshlack, and G. K. Smyth et al. Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature* 440:242–245, 2006.

[6]Gu, Z., S. A. Rifkin, K. P. White, and W. H. Li. Duplicate genes increase gene expression diversity within and between species. *Nat. Genet*. 36:577–579, 2004.

[7]Gu X. and Z. Su. Tissue-Driven Hypothesis of Genomic Evolution and Sequence-Expression Correlations. Proc. Natl. Acad. Sci. U. S. A. 2006 (Submitted).

[8]Harris, M. A., J. Clark, and A. Ireland et al. The Gene Ontology (GO) database and informatics resource. *Nucleic. Acids Res*. 32:D258–D261, 2004.

[9]Jordan, I. K., L. Marino-Ramirez, and E. V. Koonin. Evolutionary significance of gene expression divergence. *Gene* 345:119–126, 2005.

[10]Khaitovich, P., W. Enard, M. Lachmann, and S. Paabo. Evolution of primate gene expression. *Nat. Rev. Genet*. 7:693–702, 2006.

[11]Khaitovich, P., I. Hellmann, and W. Enard et al. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* 309:1850–1854, 2005.

[12]Khaitovich, P., B. Muetzel, and X. She et al. Regional patterns of gene expression in human and chimpanzee brains. *Genome Res*. 14:1462–1473, 2004.

[13]Khaitovich, P., K. Tang, and H. Franz et al. Positive selection on gene expression in the human brain. *Curr. Biol*. 16:R356–358, 2004.

[14]Khaitovich, P., G. Weiss, and M. Lachmann et al. A neutral model of transcriptome evolution. *PLoS Biol*. 2:E132.

[15]King, M. C. and A. C. Wilson. Evolution at two levels in humans and chimpanzees. *Science* 188:107–116, 1975.

[16]Lemos, B., B. R. Bettencourt, C. D. Meiklejohn, and D. L. Hartl. 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*. 22:1345–1354, 2005.

[17]Liao, B. Y. and J. Zhang. Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Mol. Biol. Evol*. 23:530–540, 2006.

[18]Nuzhdin, S. V., M. L. Wayne, K. L. Harmon, and L. M. McIntyre. Common pattern of evolution of gene expression level and protein sequence in Drosophila. *Mol. Biol. Evol*. 21:1308–1317, 2004.

[19]Rifkin, S. A., D. Houle, J. Kim, and K. P. White. A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. *Nature* 438:220–223, 2005.

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

[21]Su, A. I., M. P. Cooke, and K. A. Ching et al. Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci. U. S. A*. 99:4465–4470, 2002.

[22]Su, A. I., T. Wiltshire, and S. Batalov et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U. S. A*. 101:6062–6067, 2004.

[23]Whitehead, A. and D. L. Crawford. Variation within and among species in gene expression: raw material for evolution. *Mol. Ecol*. 15:1197–1211, 2006.

[24]Whitehead, A. and D. L. Crawford. Neutral and adaptive variation in gene expression. *Proc. Natl. Acad. Sci. U. S. A*. 103:5425–5430, 2006.

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