Gene Expression Evolves Faster in Narrowly Than in Broadly Expressed Mammalian Genes

Jing Yang,* Andrew I. Su,† and Wen-Hsiung Li*

*Department of Ecology and Evolution, University of Chicago; and †Genomics Institute of the Novartis Research Foundation, San Diego

Despite much recent interest, it remains unclear what determines the rate of evolution of gene expression. To study this issue we develop a new measure, called “Expression Conservation Index” (ECI), to quantify the degree of tissue-expression conservation between two homologous genes. Applying this measure to a large set of gene expression data from human and mouse, we show that tissue expression tends to evolve rapidly for genes that are expressed in a limited number of tissues, whereas tissue expression can be conserved for a long time for genes expressed in a large number of tissues. Therefore, expression breadth is an important determinant for evolutionary conservation of tissue expression. In addition, we find a rapid decrease in ECI with the synonymous divergence between duplicate genes, suggesting fast divergence in tissue expression between duplicate genes.

Introduction

It has been commonly thought that expression of a gene in a tissue usually implies a function of the gene in that tissue. This traditional view predicts a slow rate of evolution in tissue expression because the function of a gene would change slowly in evolutionary time. This prediction does not seem to hold in general in view of recent discoveries of incongruent expression profiles between many human and mouse orthologous genes (Huminiecki and Wolfe 2004; Yanai, Graur, and Ophir 2004). Further, it has been found that gene duplication allows rapid change in gene expression (Gu et al. 2002b; Makova and Li 2003; Huminiecki and Wolfe 2004; Gu, Zhang, and Huang 2005). However, it remains unclear what factors determine the rate of evolution of gene expression. We pursue this issue, using a recent data set that contains the expression data of a large number of human genes in 79 human tissues and a large number of mouse genes in 60 mouse tissues (Su et al. 2004). This data set allows a detailed examination of the evolution of tissue expression between human and mouse orthologous genes.

Presently, the most commonly used measure of expression pattern similarity between two genes is the Pearson correlation coefficient between the expression levels of the two genes in different tissues. We use this measure to show that gene expression profile has greatly diverged between human and mouse genes, in agreement with the results of Yanai, Graur, and Ophir (2004) and Huminiecki and Wolfe (2004). In addition, we develop a new measure that is suitable for quantifying the conservation of the expression of a gene among tissues. Using this new measure we compare the rates of expression divergence in narrowly and broadly expressed genes because it has been found that housekeeping genes evolve more slowly in protein sequence than tissue-specific genes (A. L. Hughes and M. K. Hughes 1995; Hastings 1996; Duret and Mouchiroud 2000; Zhang and Li 2004). Further, we study the rate of expression divergence between human duplicate genes.

Key words: gene expression evolution, expression conservation, duplicate genes, transcription factors.

E-mail: whli@uchicago.edu.

*Corresponding author.

Received 2004 June 23; accepted 2004 October 25

Materials and Methods

Orthologous Genes in Human and Mouse

We use the 3,055 orthologous human and mouse gene pairs that were used by Iwama and Gojobori (2004) in their analysis of the 8-kb upstream nucleotide sequences of genes. These are nuclear protein-coding genes and have the same official gene symbols for human and mouse in RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/) (Pruitt 2005). For each human and mouse gene studied, the sequence was retrieved from the Ensembl database using the EnsMart tool (http://www.ensembl.org/) (Birney et al. 2004).

Gene Expression Data

Human and mouse gene expression data were from the second version of Gene Expression Atlas, which is a compendium of gene expression experiments that surveyed expression patterns of the human and mouse transcriptomes in a panel of normal physiological tissues (Su et al. 2004). In addition to the Affymetrix HG-U133A array, this study used two custom-made arrays (GFn1H and GFn1M) for human and mouse. In total, 79 human and 60 mouse tissues were studied. (We merged the spinal cord upper and lower part in the mouse data as the homologous tissue to the spinal cord in human. Therefore, the total number of tissues studied in mouse became 60 instead of 61.) Only 30 tissues were shared by the human and mouse data sets, and they were used as homologous tissues for expression comparison (adipocyte, adrenal gland, amygdala, bone marrow, cerebellum, dorsal root ganglion, heart, hypothalamus, kidney, liver, lung, lymph node, olfactory bulb, ovary, pancreas, CD4+ Tcells, CD8+ Tcells, pituitary, placenta, prostate, salivary gland, skeletal muscle, spinal cord, testis, thymus, thyroid, tongue, trachea, trigeminal ganglion, and uterus).

The results presented here were based on data generated from applying the MAS5 condensation algorithm to the Affymetrix data; the algorithm reports an average difference (AD) value for each gene, which is an estimate of the expression level in that sample (Hubbell, Liu, and Mei 2002; Liu et al. 2002). The results were qualitatively the same when using data processed using the GC content adjusted-robust multi-array (GC-RMA) algorithm, which computes expression values from probe intensity values incorporating probe sequence information (Wu et al. 2004).
we present only the analysis with a cutoff at AD
sue. Because the conclusions were qualitatively the same,
threshold for the definition of expression of the gene in a tis-
addition, an AD value of 150 was also used as a relaxed
later analysis; 1,975 orthologous gene pairs are retained. In
human, but not in mouse for those tissues studied (or vice
though it is also possible that this gene is only expressed in
because the probe set was ‘‘dead’’ due to technical reasons,
responding human orthologous probe set had normal ex-
found that for some probe sets in mouse, the AD values
remaining probe sets for the given gene.
set, we discard all the low-confidence probe sets if higher
of gene expression. So for genes with more than one probe
Affymetrix IDs) are considered lower confidence reporters
with a higher likelihood of cross-hybridization between
hybridized to a second chip. However, because one group
of individuals could not provide a large enough amount of
RNA for all tissues, several different groups of individuals
were used for different tissues. For the human samples,
because samples were only available through commercial
or postmortem sources, less control could be applied.
Nevertheless, samples generally represent greater than
four individuals. Full details of the sample annotation
and preparation are given in the Su et al. (2004) at
http://wombat.gnf.org/.
We took the arithmetic mean of the AD values and
used it as the measure of the expression level for the
corresponding gene in a tissue. Probe sets containing probes
with a higher likelihood of cross-hybridization between
genes (indicated by a suffix of ‘‘_x_at’’ or ‘‘_s_at’’ in the
Affymetrix IDs) are considered lower confidence reporters
of gene expression. So for genes with more than one probe
set, we discard all the low-confidence probe sets if higher
confidence ones are available and take the average over the
remaining probe sets for the given gene.
In this study, we use an AD value of 200 as the thresh-
hold for calling a gene ‘‘expressed in a given tissue’’ (Su et al.
2002). However, upon closer inspection of the data, we
found that for some probe sets in mouse, the AD values
were all well below 200 across the 60 tissues, while its cor-
responding human orthologous probe set had normal ex-
pression in several tissues, and vice versa. This can be
because the probe set was ‘‘dead’’ due to technical reasons,
though it is also possible that this gene is only expressed in
human, but not in mouse for those tissues studied (or vice
versa). For simplicity, we discarded such probe sets in our
later analysis; 1,975 orthologous gene pairs are retained. In
addition, an AD value of 150 was also used as a relaxed
threshold for the definition of expression of the gene in a tis-
ue. Because the conclusions were qualitatively the same,
we present only the analysis with a cutoff at AD = 200.

Intra- and Interspecies Variation in Expression Level
We compared inter- and intraspecies variation in
expression level. In the data we used, only two experimen-
tal replicates (samples) for each tissue were obtained in
each species and because one group of individuals could
not provide a large enough amount of RNA for all tis-
sues, several different groups of individuals were used
for different tissues. Because of these limitations, we
cannot calculate the within-species (among individuals)
variation in the standard way. However, we show below
that within-species variation is small relatively to the
between-species variation.
Let us use the human data as an example. For each of
the 1,975 genes used in our study, we first compute the
within-species variation. For each human tissue, we obtain
the absolute value of the difference between the two expres-
sion values. In this manner, we obtain 30 such values for
the 30 tissues. Second, we compute the between-species
differences. For each tissue, we obtain the average expres-
sion value in human, the average expression value in
mouse, and then the absolute value of the difference be-
tween the two values. For the 30 tissues we obtain 30 such
values that represent the between-species variation. Then
for each gene, we use the t-test to test whether the 30
between-species differences are significantly greater than
the 30 within-species replicate differences. Indeed, all tests
(all 1,975 genes) are significant. The same conclusion holds
for the mouse data. So, we can conclude that the between-
species variation is in general significantly larger than the
within-species variation in both human and mouse.

Measures of Expression Similarity
We consider two measures of expression similarity
between genes. The first one is the Pearson correlation
coefficient (r) between the AD values of the human and
orthologous genes. Because when the AD value is below 200 (or 150) r mainly reflects background noise,
in computing the r value for a gene we exclude all tissues
that have an AD value below 200 (or 150) in both species.
Further, to have a sufficiently large number of points for
computing r, we keep only the pairs of human and mouse
genes for which the gene is expressed in at least 5 of the 30
tissues (AD value ≥200) in one or both species. We have
also calculated the Spearman’s rank correlation, which is
less likely to be affected by extreme values compared to
Pearson’s, and come to qualitatively the same conclusion.
Second, we develop a new measure, called the expres-
sion conservation index (ECI) between two species. A gene
is said to have a conserved expression in a tissue if it is
expressed in that tissue in both species but a divergent ex-
pression in a tissue if it is expressed in only one of the two
species but not in both. For a gene under study, let n be the
number of tissues with a conserved expression and N be the
average of the number (N1) of tissues in which this gene is
expressed in human and the number (N2) of tissues in which
this gene is expressed in mouse. Then the ECI for the gene
is defined as (n + 0.5)/(N + 0.5); we add 0.5 to both the
numerator and the denominator to reduce the effect of a
small N. In this formulation, we use N = (N1 + N2)/2
to estimate the number of tissues in which this gene showed
expression in the common ancestor of the two species under
study, assuming that the number of tissues in which the gene
 gained new expression is equal to the number of tissues in
which the gene lost expression. That is, we assume an equi-
librium condition under which the number of tissues in
which a gene lost expression is equal to the number of tis-
sues in which the gene gained expression during the time
period under study. Thus, ECI is intended to estimate the
proportion of tissue expressions that have been conserved
since the divergence of the two species or duplicate genes.
This formulation is similar to the formulation of Nei and Li
(1979) for the evolution of restriction sites in DNA sequen-
ces. In addition, we consider only the gene pairs in which at
least one member of the pair is expressed in at least 2 of the
30 tissues studied in both human and mouse (i.e., N ≥ 1).

N = (N1 + N2)/2
Duplicate Gene Identification

We used the method of Gu et al. (2002a) to identify the duplicated gene pairs in the human genome and the PAML package with default parameters (Yang et al. 1997) to estimate $K_s$ and $K_a$, which are the numbers of substitutions per synonymous and nonsynonymous site, respectively. We choose only duplicate gene pairs with $K_s \leq 0.4$ as the human lineage-specific duplicate genes.

Results

Low Correlation in Expression Level Between Human and Mouse Genes

We first consider the correlation ($r$) in expression level between human and mouse orthologous genes. We use a set of well-defined human and mouse orthologous genes studied by Iwama and Gojobori (2004). However, we exclude genes that are expressed in fewer than 5 of the 30 tissues studied in both human and mouse because when the number of data points is small the computed $r$ may be heavily affected by a single point. Figure 1a, which is based on the AD values, reveals a peak near 0 in the distribution of $r$ values and a large proportion (>70%) of gene pairs with $r < 0.5$. Therefore, many human and mouse orthologous genes appear to have diverged in expression to the extent as two unrelated genes. This observation is in agreement with the results of Yanai, Graur, and Ophir (2004) and Huminiecki and Wolfe (2004), who used the first version of the Gene Expression Atlas (Su et al. 2002), which is considerably less extensive than the current version. All these results imply rapid evolution of expression profile in many mammalian genes. When we use the GC-RMA data instead of the AD values, the same pattern holds, but the distribution is even more centered at 0 (fig. 1b) compared to the more spread distribution in figure 1a.

Expression Breadth versus Expression Conservation

We define a gene to be broadly expressed if it is expressed in $\geq 30$ of the 79 tissues studied in human and to be non–broadly expressed if otherwise; we consider the human data because more human tissues have been studied than mouse tissues. Figure 2a shows the $ECI$ distributions for broadly expressed and non–broadly expressed genes. Note that most non–broadly expressed genes have an $ECI$ value $< 0.5$; that is, they have diverged in expression in over half of the tissues compared. In contrast, only 50% of the broadly expressed genes have conserved gene expression in over half of the tissues compared. In conclusion, tissue expression evolves faster in narrowly expressed genes than in broadly expressed genes. This conclusion still holds if the definition of expression of a gene in a tissue is relaxed to $AD \geq 150$ (fig. 2b) or when we use the GC-RMA data.

The importance of expression breadth as a determinant of expression conservation is further supported by the following analysis. We use the 49 tissues studied in human but not in mouse to define the expression breadth of a gene. We count the number of tissues in which a gene is expressed in these 49 tissues and divide the expression levels of orthologous human and mouse genes. The expression values are the AD values from Gene Expression Atlas (Su et al. 2004).
breadth into 10 bins each of the 5 tissues. We then take the average of the ECI values for the genes in each bin, which are computed from the 30 tissues studied in both human and mouse and plot the value against expression breadth (fig. 3). It is seen that in general the average ECI value increases with the expression breadth. When the 30 tissues studied in mouse but not in human are used to define the expression breadth, a similar pattern is also found (fig. 3). All of the above analyses include duplicate genes. However, exclusion of genes that have been duplicated after the human-mouse split does not qualitatively affect the above conclusions.

Are Transcription Factor Genes More Conservative in Tissue Expression?

Iwama and Gojobori (2004) have recently found that the 8-kb upstream region of a gene tends to be much better conserved in transcription factor (TF) genes than in non-TF genes. On the basis of this observation one may hypothesize that gene expression evolves faster in non-TF genes than in TF genes. However, figure 4 suggests otherwise. Further, no clear correlation between the degree of conservation in the 8-kb region of a gene and its tissue expression conservation was found in our analysis (data not shown). Of course, this may not necessarily imply that there is no relationship between conservation of 5′ regulatory sequences of genes and expression conservation but may imply that sequence specificity in the upstream 8-kb region of a gene is loose or only small subregions of the 8-kb regions are involved in gene regulation. We note that TF-binding sites are usually only 5–15 nt long and the sequences can be degenerate, so they may not contribute strongly to the overall conservation of the 8-kb upstream region of a gene.

Duplicate Genes

We also study the correlation between Ks and ECI for human duplicated genes. We obtain a total of 114 pairs of duplicated genes in human with expression data for both genes of the pair. The Ks values are between 0.05 and 0.35; we exclude pairs with a Ks < 0.05 to reduce the effect of cross-hybridization in microarrays and also pairs with Ks > 0.35 because there are too few of them for a bin width of Ks = 0.05. We group the genes based on the Ks values with a 0.05 increment. Then we look at the relationship between the average ECI among 79 human tissues for each group and the related Ks value. We require the pair to have expression in at least two tissues to be considered for ECI. The regression result is shown in figure 5 with $R^2 = 0.86$, $P$ value of 0.007, and the slope is $-0.78$. Therefore, there is a significant negative correlation between ECI and Ks. Because a smaller ECI implies more divergent tissue expressions, our analysis shows that among human paralogous pairs, the change in tissue expression increases with the synonymous divergence or with the evolutionary time.

Discussion

The two measures of gene expression similarity used in this study have their strengths and weaknesses. This can be illustrated by the two cases in figure 6. In figure 6a, the $r$ value (0.88) is fairly high, despite the fact that, under the expression cut off point of AD = 200, the gene was expressed in as many as 24 tissues in human but in only 1 tissue in mouse among the 30 tissues studied in both human and mouse. This case clearly shows that $r$ can be strongly affected by a single tissue that happens to express the gene at a level much higher than the other tissues in both species. In comparison, the ECI value (0.12) is low, correctly reflecting the expression divergence between the two species. On the other hand, in figure 6b, the ECI value (0.98) is very...
high because the gene was expressed in most of the 30 tissues in both human (30 tissues) and mouse (28 tissues), while the \( r \) value is low (–0.14) because the differences in expression level between the two species fluctuated greatly over the 30 tissues. In this case, the \( ECI \) does not reflect well the absence of correlation in gene expression level between the two species among tissues. However, we would argue that the most important question in the study of gene expression is whether the gene is expressed in a given tissue or not, while the level of expression is of secondary importance; from this point of view, the high ECI in figure 6b is indeed a good expression indicator. For this reason, \( ECI \) may be a better measure of gene expression similarity than \( r \). Of course, the two measures are complementary, and both should be used. Moreover, \( ECI \) may be more strongly affected by measurement or experimental errors when the expression level in a tissue is close to the cut off threshold used to define tissue expression.

We have seen that many genes have a low \( ECI \) value and thus a high rate of loss of expression in a tissue or gain of expression in a new tissue. This observation suggests that in many cases the expression of a gene in a tissue may be transient and not evolutionarily stable. A possible reason for a higher conservation of tissue expression for broadly expressed genes might be because they tend to behave like housekeeping genes and so tend to be essential to the organism. This is in agreement with the observation that housekeeping genes in general tend to have a lower rate of nonsynonymous substitution than tissue-specific genes (A. L. Hughes and M. K. Hughes 1995; Hastings 1996; Duret and Mouchiroud 2000; Zhang and Li 2004). On the other hand, for narrowly expressed genes, expression in a tissue may not be essential and therefore the expression may become lost in evolution. Large-scale gene expression studies in mammals suggest that there can often be leaky (unneeded) expression in noncoding regions (Kapranov et al. 2002; Johnson et al. 2005) and this may also be true for coding regions. Thus, it is possible that the expression of one member of an orthologous pair in a tissue is accidental and may not be truly functional. This situation may be more often for narrowly expressed genes than for broadly expressed genes. Of course, it is also possible that the expression level of a tissue-specific gene is more dependent on the developmental stage or physiological conditions of the subject and this can increase measurement errors and lower the \( ECI \) value.

Further, one may argue that widely expressed genes may tend to have a higher \( ECI \) than narrowly expressed genes because for a gene that has already been expressed in many of the 30 homologous tissues, a new tissue expression in human and a new tissue expression in mouse should have a higher chance to be in the same tissue than a gene that has been expressed in only a few tissues. However, this possibility can at best be only part of the reason for the higher \( ECI \) for broadly expressed genes for two reasons. First, the breadth of gene expression in figure 3 was defined using the tissues that were studied only in human but not in mouse, so that the definition was independent of the 30 homologous tissues used to study the \( ECI \). We also note that the expression breadth used in figure 2 was defined using the 79 human tissues without any regard of the tissues studied in mouse. Second, let us consider the following analysis. For the broadly and non–broadly expressed gene groups, we select those genes that have the same number of tissue expressions in the 30 tissues studied in both human and mouse. In this way, the \( ECI \) value is not affected by the expression breadth in the 30 homologous tissues because among the 30 homologous tissues the numbers of tissues that express the non–broadly and broadly expressed genes are equal. In total, there are 158 such gene pairs (table 1). Note that 56% of these pairs have a higher \( ECI \) in broadly expressed genes, which is significantly higher than the corresponding proportion (30%) for non–broadly expressed genes, supporting our conclusion.

Because the function of a tissue-specific gene is usually highly specific, one may argue that its function and expression are expected to have a high degree of conservation among different species, but this expectation is

### Table 1

Comparison of \( ECI \) Values for Broadly and Non–Broadly Expressed Genes When the Number of Tissues Expressed in the 30 Human and Mouse Homologous Tissues Are the Same

<table>
<thead>
<tr>
<th>( ECI ) Comparison</th>
<th>Number of Gene Pairs</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>( ECI ) broadly &gt; ( ECI ) nonbroadly</td>
<td>88</td>
<td>0.56</td>
</tr>
<tr>
<td>( ECI ) broadly &lt; ( ECI ) nonbroadly</td>
<td>47</td>
<td>0.30</td>
</tr>
<tr>
<td>( ECI ) broadly = ( ECI ) nonbroadly</td>
<td>23</td>
<td>0.14</td>
</tr>
<tr>
<td>Total</td>
<td>158</td>
<td>1</td>
</tr>
</tbody>
</table>
not supported by our study. First, we note that defining tissue-specific genes is not simple. As microarray data contain much noise, it is difficult to find a consensus threshold to define whether a gene is expressed in a tissue or not. Second, tissue-specific genes may not be truly tissue specific under different physiological conditions. Finally, even if we neglect the above two assumptions and just set up one single cut off point for expression in a tissue, we find that among the orthologous genes under study there are 90 single-tissue expression genes in human and 226 single-tissue expression genes in mouse. Surprisingly, these two sets of genes share only 18 genes in common and only 6 out of the 18 genes have a conserved tissue expression pattern (they are expressed in the same single tissue in both species). Therefore, this observation suggests that tissue-specific genes actually tend to evolve fast in expression pattern.

It has been proposed that neutral evolution, i.e., evolution by mutation and random drift, of gene expression is widespread because there was no clear correlation between sequence divergence in coding regions and expression divergence and because incongruent expression profiles were found between human and chimpanzee and between human and mouse orthologous genes (Khaitovich et al. 2004; Yanai, Graur, and Ophir 2004). This is also observed in our study when we use the Pearson correlation coefficient as the measure of expression similarity. However, a consideration of the tissue distribution of gene expression suggests that the breadth of gene expression is an important determinant for the conservation of gene expression and broadly expressed genes may show a high degree of conservation in tissue expression.

Acknowledgments

We thank Justin Borevitz, Jake Byrnes, Jianying Gu, Geoffrey Morris, and Shin-Han Shiu for discussions and suggestions. We also thank the reviewers for valuable suggestions. This study was supported by National Institutes of Health grants.

Literature Cited


Douglas Crawford, Associate Editor

Accepted June 24, 2005