DOI: 10.1534/genetics.109.110247

Note

A Problem With the Correlation Coefficient as a Measure of Gene Expression Divergence

Vini Pereira,[1] David Waxman and Adam Eyre-Walker

Centre for the Study of Evolution, School of Life Sciences, University of Sussex, Brighton BN1 9QG, United Kingdom

Manuscript received September 24, 2009
Accepted for publication September 25, 2009

ABSTRACT

The correlation coefficient is commonly used as a measure of the divergence of gene expression profiles between different species. Here we point out a potential problem with this statistic: if measurement error is large relative to the differences in expression, the correlation coefficient will tend to show high divergence for genes that have relatively uniform levels of expression across tissues or time points. We show that genes with a conserved uniform pattern of expression have significantly higher levels of expression divergence, when measured using the correlation coefficient, than other genes, in a data set from mouse, rat, and human. We also show that the Euclidean distance yields low estimates of expression divergence for genes with a conserved uniform pattern of expression.

IT is now possible to measure the expression levels of thousands of genes in multiple tissues at multiple times. This has led to investigations into the evolution of gene expression and how the pattern of expression changes on a genomic scale. In some analyses, the evolution of expression is considered only within one tissue, but in many studies the evolution across multiple tissues is investigated. In this latter case, the evolution of an expression profile—a vector of expression levels of a gene across several tissues—is considered.

Several different statistics have been proposed to measure the divergence between gene expression profiles. The two most popular measures are the Euclidean distance (Jordan *et al.* 2005; Kim *et al.* 2006; Yanai *et al.* 2006; Urrutia *et al.* 2008) and Pearson's correlation coefficient (Makova and Li 2003; Huminiecki and Wolfe 2004; Yang *et al.* 2005; Kim *et al.* 2006; Liao and Zhang 2006a,b; Xing *et al.* 2007; Urrutia *et al.* 2008). The correlation coefficient is often subtracted from one, so that the statistic varies from zero, when there has been no expression divergence, to a maximum of two; we refer to this statistic as the *Pearson distance*. Here we describe a significant shortcoming of the Pearson distance that is not shared by the Euclidean distance.

To investigate properties of these two measures of expression divergence, we compiled a data set of 2859 orthologous genes from human, mouse, and rat for which we had microarray expression data from nine homologous tissues: bone marrow, heart, kidney, large intestine, pituitary, skeletal muscle, small intestine, spleen, and thymus). The expression data for rat came from Walker *et al.* (2004), the mouse data from Su *et al.* (2004), and the human data from Ge *et al.* (2005). Each tissue experiment had two replicates in mouse, a varying number of replicates in rat, and one in humans; some genes were also matched by multiple probe sets. To obtain an average across experiments and probe sets we processed the data as follows:

i. Raw CEL files of gene expression levels were obtained from the NCBI Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/projects/geo/).
ii. The results from the mouse, rat, and human arrays were normalized separately using both the MAS5 (Affymetrix 2001) and the RMA algorithms (Irizarry *et al.* 2003) as implemented in Bioconductor (Gentleman *et al.* 2004). The results are qualitatively similar for the two normalization procedures, although recent analyses suggest that MAS5 normalization is generally better (Ploner *et al.* 2005; Lim *et al.* 2007).
iii. The expression of each gene within a tissue was averaged across experiments and probe sets.

Supporting information is available online at http://www.genetics.org/cgi/content/full/genetics.109.110247/DC1.

[1]*Corresponding author:* Theoretical Systems Biology, Institute of Food Research, Norwich Research Park, Colney, Norwich NR4 7UA, United Kingdom. E-mail: vini.pereira@bbsrc.ac.uk

We computed expression distances (ED) between orthologous gene expression profiles, for each of the three species comparisons, rat–mouse, rat–human, and mouse–human, according to the two different distance metrics, the Euclidean distance and the Pearson distance:

$$\begin{aligned} \text{EucD} &= \sqrt{\sum_{j=1}^{k}(x_{1j}-x_{2j})^2} \\ \text{PeaD} &= 1-\frac{\sum_{j=1}^{k}(x_{1j}-\overline{x_1})(x_{2j}-\overline{x_2})}{\sqrt{\sum_{j=1}^{k}(x_{1j}-\overline{x_1})^2\sum_{j=1}^{k}(x_{2j}-\overline{x_2})^2}}. \end{aligned} \quad (1)$$

Here x_{ij} is the expression level of the gene under consideration in species i in tissue j, and $\overline{x}_i$ is the average expression level of the gene in species i across tissues. Expression levels are known in a total of k tissues.

Because expression levels are measured on different microarray platforms in the three species, we compute *relative abundance* (RA) values, before calculating the Euclidean distance (Liao and Zhang 2006a). The RA is the expression of a gene in a particular tissue divided by the sum of the expression values of that gene across all tissues. We calculated RA values to remove "probe" effects (the tendency for a gene to bind its probe set on one platform more efficiently than on another platform). Because of probe effects it is not easy to distinguish absolute changes in expression and differences in binding efficiency. Calculating RA values removes this problem from the Euclidean distance. Pearson's distance does not change under such a rescaling and so this is unnecessary.

In some analyses the logarithm of the expression or RA values are used (*e.g.*, Makova and Li 2003; Kim *et al.* 2006; Xing *et al.* 2007), and in others the expression values are used without this transformation (*e.g.*, Huminiecki and Wolfe 2004; Jordan *et al.* 2005; Yang *et al.* 2005; Liao and Zhang 2006a,b; Yanai *et al.* 2006; Urrutia *et al.* 2008). We calculated both the Pearson and the Euclidean distances on log-transformed and untransformed expression values. The results are qualitatively similar so here we present only the results obtained using the logarithm of the expression or RA values.

It is natural to expect the two measures of expression divergence to be positively correlated with one another; however, the Euclidean and Pearson distances are almost completely uncorrelated (MAS5 normalization, mouse–rat correlation coefficient = 0.06, human–rat r = 0.13, human–mouse r = 0.10; RMA normalization, mouse–rat correlation coefficient = −0.12, human–rat r = −0.00, human–mouse r = −0.08; Figure 1). This could, plausibly, be because the two statistics measure different aspects of divergence. However, irrespective of this, there is a potential problem associated with the Pearson distance. Imagine that we have a gene that is expressed at *identical* levels in all tissues in two species (*i.e.*, expression levels are uniform between tissues and

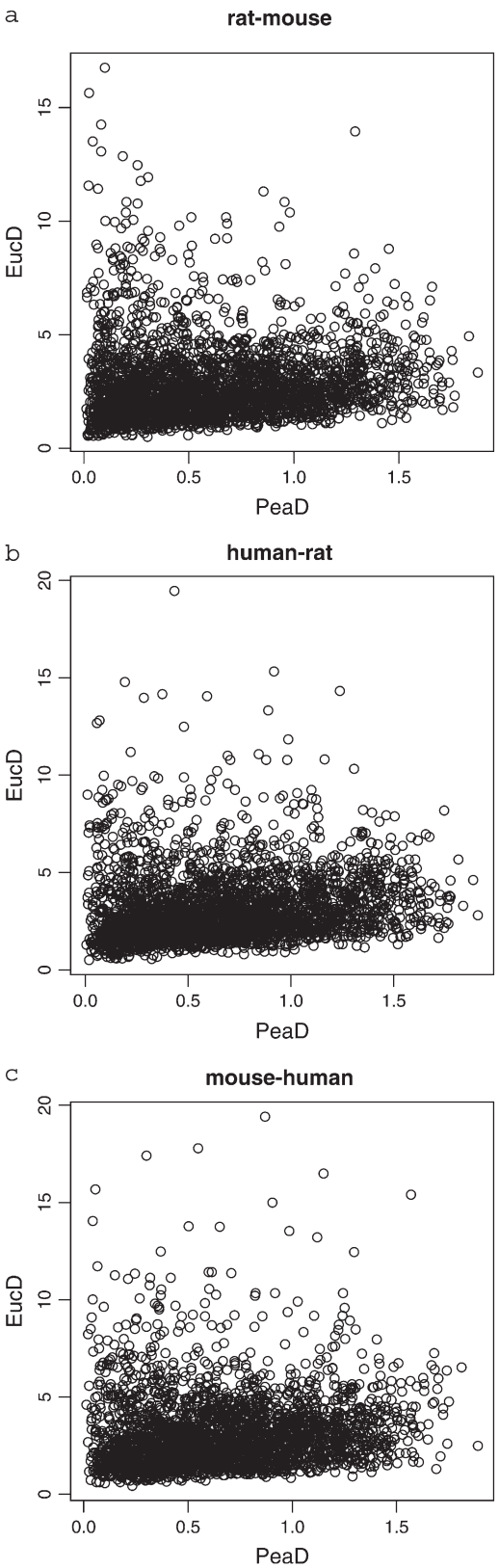

Figure 1.—The correlation between the Euclidean and Pearson distances for (a) mouse–rat, (b) human–rat, and (c) human–mouse. Only the results from MAS5 normalization are shown; qualitatively similar results were obtained with RMA.

TABLE 1

The median expression divergence for genes that have a conserved uniform pattern of expression (upper quartile of mean entropy values) *vs.* all other genes

Data set	Statistic	Conserved uniform genes	Other genes	Wilcoxon test *P*-value
MAS5 normalization				
Mouse–rat	Euclidean	1.66	2.79	$<10^{-15}$
	Pearson	0.70	0.47	$<10^{-15}$
Human–mouse	Euclidean	1.67	3.13	$<10^{-15}$
	Pearson	0.78	0.58	$<10^{-15}$
Human–rat	Euclidean	1.83	3.21	$<10^{-15}$
	Pearson	0.78	0.58	$<10^{-15}$
RMA normalization				
Mouse–rat	Euclidean	0.59	1.40	$<10^{-15}$
	Pearson	0.82	0.38	$<10^{-15}$
Human–mouse	Euclidean	0.59	1.58	$<10^{-15}$
	Pearson	0.81	0.48	$<10^{-15}$
Human–rat	Euclidean	0.58	1.55	$<10^{-15}$
	Pearson	0.73	0.50	$<10^{-15}$

also between species). We quite reasonably assume that *measured* expression levels contain noise. Thus each *measured* expression level (x_{ij}) is the sum of the (assumed) uniform expression level and an independent random number representing noise. In this case there is no real divergence in the expression profile between the species. However, the two measures of divergence may differ greatly in this case. The Euclidean distance reflects only the noise present in the data and hence will be small if the noise is small. By contrast, the Pearson distance will have a value close to 1 since the second term in PeaD in Equation 1 will be close to zero, reflecting the fact that the noise components of different expression levels are independent. Thus the Pearson distance will give the impression that expression divergence is great, but all this apparent divergence is noise. This will be a problem with Pearson's distance whenever measurement error is of the same magnitude as the differences in expression between tissues. This will therefore tend to be a problem for lowly expressed genes, where measurement error can be large relative to the true value.

The above example is unrealistic because real gene expression profiles are rarely perfectly uniform. To investigate whether this shortcoming of the Pearson distance is a problem in real data sets, we determined genes with a relatively uniform pattern of expression in all three species considered above. To do this we computed the *entropy* of a gene's expression, which is a measure of uniformity in expression across tissues (Schug *et al.* 2005): the higher the value of the entropy, the more uniform is the expression. We calculated the entropy for each gene in each of the three species, averaged these across species, and then took those genes in the upper quartile of mean entropy values as a data set of genes with a relatively conserved pattern of uniform expression.

It is natural to expect those genes with a conserved uniform pattern of expression to have relatively low expression divergence; however, on average these genes have significantly higher Pearson distances than other genes (Table 1; Figure 2; supporting information, Figure S1 and Figure S2). By contrast, the Euclidean

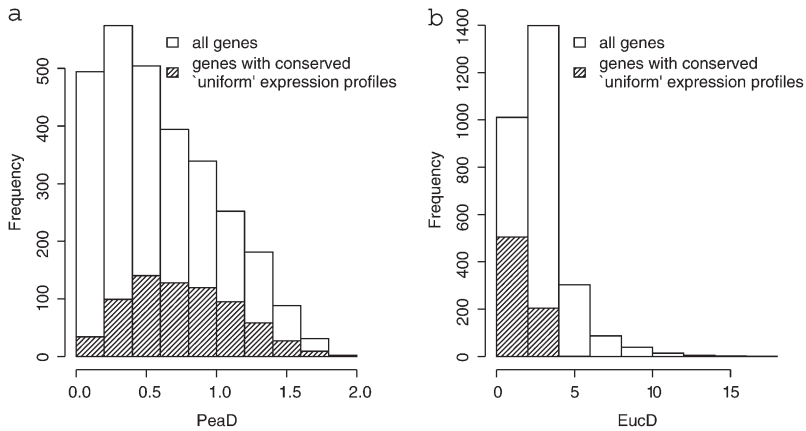

Figure 2.—The distribution of expression divergence values for those genes with a uniform pattern of expression that is conserved across species *vs.* the distribution for all genes for (a) Pearson and (b) Euclidean distances for mouse–rat. We present similar values for human–mouse and human–rat in Figure S1 and Figure S2. Only the results from MAS5 normalization are shown; qualitatively similar results were obtained with RMA.

distance shows the pattern one would anticipate; all of the conserved uniform genes have low expression divergence. It therefore seems likely that the Pearson distance is sensitive to measurement error and hence may not be a good measure of expression divergence.

We note that there are two additional advantages of the Euclidean distance. First, it can take into account differences in the absolute level of expression if those data are available, either because the method of assay allows this, for example, if ESTs, SAGE, sequencing, or RNA-Seq data are used, or because expression in the two species has been assessed on the same platform using probes that are conserved between the two species. Second, the square of the Euclidean distance is expected to increase linearly with time. KHAITOVICH *et al.* (2004) have previously shown that the squared difference in log expression level increases linearly with time under a Brownian motion model of gene expression evolution. It is therefore expected that the squared Euclidean distance will increase with time since the squared Euclidean distance is the sum of the squared differences across tissues. We prove this in File S1; we also show that this linearity holds, approximately, when relative abundance values are used (see also PEREIRA *et al.* 2009).

We are grateful to a referee for helpful comments. V.P. and A.E.W. were supported by the Biotechnology and Biological Sciences Research Council.

LITERATURE CITED

AFFYMETRIX, 2001 *Statistical Algorithms Reference Guide.* Affymetrix, Santa Clara, CA.

GE, X., S. YAMAMOTO, S. TSUTSUMI, Y. MIDORIKAWA, S. IHARA *et al.*, 2005 Interpreting expression profiles of cancers by genome wide survey of breadth of expression in normal tissues. Genomics **86:** 127–141.

GENTLEMAN, R. C., V. J. CAREY, D. M. BATES, B. BOLSTAD, M. DETTLING *et al.*, 2004 Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. **5:** R80.

HUMINIECKI, L., and K. H. WOLFE, 2004 Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. Genome Res. **14:** 1870–1879.

IRIZARRY, R. A., B. HOBBS, F. COLLIN, Y. D. BEAZER-BARCLAY, K. J. ANTONELLIS *et al.*, 2003 Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics **4:** 249–264.

JORDAN, I. K., L. MARINO-RAMIREZ and E. V. KOONIN, 2005 Evolutionary significance of gene expression divergence. Gene **345:** 119–126.

KHAITOVICH, P., G. WEISS, M. LACHMANN, I. HELLMANN, W. ENARD *et al.*, 2004 A neutral model of transcriptome evolution. PLoS Biol. **2:** E132.

KIM, R. S., H. JI and W. H. WONG, 2006 An improved distance measure between the expression profiles linking co-expression and co-regulation in mouse. BMC Bioinformatics **7:** 44.

LIAO, B.-Y., and J. ZHANG, 2006a Evolutionary conservation of expression profiles between human and mouse orthologous genes. Mol. Biol. Evol. **23:** 530–540.

LIAO, B. Y., and J. ZHANG, 2006b Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution. Mol. Biol. Evol. **23:** 1119–1128.

LIM, W. K., K. WANG, C. LEFEBVRE and A. CALIFANO, 2007 Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. Bioinformatics **23:** i282–i288.

MAKOVA, K. D., and W. H. LI, 2003 Divergence in the spatial pattern of gene expression between human duplicate genes. Genome Res. **13:** 1638–1645.

PEREIRA, V., D. ENARD and A. EYRE-WALKER, 2009 The effect of transposable element insertions on gene expression evolution in rodents. PLoS ONE **4:** e4321.

PLONER, A., L. D. MILLER, P. HALL, J. BERGH and Y. PAWITAN, 2005 Correlation test to assess low-level processing of high-density oligonucleotide microarray data. BMC Bioinformatics **6:** 80.

SCHUG, J., W. P. SCHULLER, C. KAPPEN, J. M. SALBAUM, M. BUCAN *et al.*, 2005 Promoter features related to tissue specificity as measured by Shannon entropy. Genome Biol. **6:** R33.

SU, A. I., T. WILTSHIRE, S. BATALOV, H. LAPP, K. A. CHING *et al.*, 2004 A gene atlas of the mouse and human protein coding transcriptomes. Proc. Natl. Acad. Sci. USA **101:** 6062–6067.

URRUTIA, A. O., L. B. OCANA and L. D. HURST, 2008 Do Alu repeats drive the evolution of the primate transcriptome? Genome Biol. **9:** R25.

WALKER, J. R., A. I. SU, D. W. SELF, J. B. HOGENESCH, H. LAPP *et al.*, 2004 Applications of a rat multiple tissue gene expression data set. Genome Res. **14:** 742–749.

XING, Y., Z. OUYANG, K. KAPUR, M. P. SCOTT and W. H. WONG, 2007 Assessing the conservation of mammalian gene expression using high-density exon arrays. Mol. Biol. Evol. **24:** 1283–1285.

YANAI, I., J. O. KORBEL, S. BOUE, S. K. MCWEENEY, P. BORK *et al.*, 2006 Similar gene expression profiles do not imply similar tissue functions. Trends Genet. **22:** 132–138.

YANG, J., A. I. SU and W. H. LI, 2005 Gene expression evolves faster in narrowly than in broadly expressed mammalian genes. Mol. Biol. Evol. **22:** 2113–2118.

Communicating editor: I. HOESCHELE

GENETICS

Supporting Information
http://www.genetics.org/cgi/content/full/genetics.109.110247/DC1

A Problem With the Correlation Coefficient as a Measure of Gene Expression Divergence

Vini Pereira, David Waxman and Adam Eyre-Walker

DOI: 10.1534/genetics.109.110247

FILE S1

A problem with the correlation coefficient as a measure of gene expression divergence.

Vini Pereira, David Waxman & Adam Eyre-Walker

In this supplementary information we establish two results under a random walk model of gene expression from an ancestral state. (i) The square of the Euclidean distance of gene expression profiles increase *linearly with time*. (ii) The Euclidean distance of the relative abundance values of a gene *approximately* increases *linearly with time*. (The relative abundance value of a gene in a tissue is its expression level in that tissue divided by the sum of its expression values over all tissues).

We consider evolution of gene expression levels of a number of different genes in k tissues. Let x_j denote the expression level of a particular gene, in a particular species, in tissue j. We collect these expression levels into a vector $(x_1, x_2, ..., x_k)$ representing the expression level of the gene in all k tissues and refer to this vector as the expression level profile of the gene.

We make the assumption that the expression levels of different genes, in different tissues, are the outcome of *independent multiplicative random walks* from an ancestral expression level. The ancestor is taken to occur at time $t = 0$ and the expression level of a particular gene in tissue j in the ancestor is written $x_j(0)$. The corresponding gene expression level after t generations is $x_j(t) = f_t f_{t-1} ... f_2 f_1 x_j(0)$ where the factors f_k are independent random variables, where $\log f_k$ is normal with mean 0 and variance σ_j (i.e., the f_k are log-normal random variables). The vanishing mean of $\log f_k$ corresponds to f_k and f_k^{-1} being equally likely, so genes are equally likely to be up or down regulated by the same factor. We explicitly allow the variance of $\log f_k$ to depend on tissue type (i.e., on j).

Since a sum of independent normal random variables is also normal, it follows that $\log x_j(t)$ is a normal random variable with mean $\log x_j(0)$ and variance $\sigma_j^2 t$ and we can write $\log x_j(t) = \log x_j(0) + \sigma_j \sqrt{t} Z_j$ or equivalently $x_j(t) = x_j(0) \exp\left(\sigma_j \sqrt{t} Z_j\right)$, where here and elsewhere, Z's with different *labels* are independent and identically distributed normal random variables with mean zero and variance unity.

Note that the ancestral expression values (the $x_j(0)$), for different genes and different tissues generally take different values.

Results for unscaled expression profiles

We consider the expression level of a given gene in different tissues. The log expression-level profile of the gene is

$$\mathbf{L}(t) = \ell + \sqrt{t}(\sigma_1 Z_1, \sigma_2 Z_2, ..., \sigma_k Z_k)$$

Here $\ell = (\log x_1(0), \log x_2(0), ..., \log x_k(0))$ is the log expression-level of the gene in the common ancester which exists at time $t = 0$. The divergence of expression level profile from the ancestral value of the gene is $\mathbf{L}(t) - \ell = \sqrt{t}(\sigma_1 Z_1, \sigma_2 Z_2, ..., \sigma_k Z_k)$ and this has an expected value of zero: $E\left[\mathbf{L}(t) - \ell\right] = 0$. The "branch length" associated with the expression level of the gene is the squared Euclidean length $\left\|\mathbf{L}(t) - \ell\right\|^2 = t\sum_{j=1}^{k}\left(\sigma_j\right)^2\left(Z_j\right)^2$. This evidently increases linearly with time, as does its expected value: $E[\left\|\mathbf{L}(t) - \ell\right\|^2] = t\sum_{j=1}^{k}\left(\sigma_j\right)^2$.

Scaled profiles

Let us now consider scaled data. The data is scaled (normalised) such that for a given gene, the *sum of the normalised expression levels over all tissues is unity*. The normalised expression profile for a given gene is thus

$$\gamma_j(t) = \frac{x_j(t)}{\sum_{l=1}^{k} x_l(t)} \tag{1}$$

and this satisfies $\sum_{j=1}^{k}\gamma_j(t) = 1$. In terms of the Z's, we have

$$\gamma_j(t) = \frac{x_j(0)\exp\left(\sigma_j\sqrt{t}Z_j\right)}{\sum_{l=1}^{k} x_l(0)\exp\left(\sigma_l\sqrt{t}Z_l\right)}. \tag{2}$$

Taking logs yields

$$L_j(t) \stackrel{\text{def}}{\equiv} \ln\left[\gamma_j(t)\right] = \sqrt{t}\sigma_j Z_j + \ell_j(t)$$

where

$$\ell_j(t) \stackrel{\text{def}}{\equiv} \ln\left[\gamma_j(0)\right] - \ln\left[\sum_{l=1}^{k}\gamma_l(0)\exp\left(\sigma_l\sqrt{t}Z_l\right)\right]. \tag{3}$$

We take the $L_j(t)$ for a given gene to constitute elements of a *row* vector $\mathbf{L}$ and similarly the $\ell_j(t)$ constitute elements of a *row* vector $\ell(t)$. Then

$$\mathbf{L}(t) = \sqrt{t}(\sigma_1 Z_1, \sigma_2 Z_2, ..., \sigma_k Z_k) + \ell(t) \tag{4}$$

is the log transformed, normalised expression profile of a given gene at time t. The *change* in this quantity from the ancestral value is

$$\mathbf{L}(t) - \mathbf{L}(0) = \sqrt{t}(\sigma_1 Z_1, \sigma_2 Z_2, ..., \sigma_k Z_k) + \ell(t) - \ell(0) \tag{5}$$

The last term on the right hand side of Eq. (3) complicates matters. We make an approximation, assuming that for all tissues

$$\sigma_j^2 t \ll 1. \tag{6}$$

This condition corresponds to changes in expression levels between extant species and the common ancestor being typically small. Then

$$\ell_j(t) \simeq \ln\left[\gamma_j(0)\right] - \ln\left[\sum\nolimits_{l=1}^{k} \gamma_l(0)\left(1+\sigma_l\sqrt{t}Z_l\right)\right]$$

$$= \ln\left[\gamma_j(0)\right] - \ln\left(1+\sum\nolimits_{l=1}^{k} \gamma_l(0)\sigma_l\sqrt{t}Z_l\right)$$

$$\simeq \ln\left[\gamma_j(0)\right] - \sqrt{t}\sum\nolimits_{l=1}^{k} \gamma_l(0)\sigma_l Z_l \tag{7}$$

with corrections of order $\sigma^2 t$. Then

$$L_j(t) - L_j(0) \simeq \sqrt{t}\left(\sigma_j Z_j - \sum\nolimits_{l=1}^{k} \gamma_l(0)\sigma_l Z_l\right). \tag{8}$$

This has an expected value that vanishes to leading order in $\sqrt{\sigma^2 t}$ and a squared Euclidean length of $\|\mathbf{L}(t)-\mathbf{L}(0)\|^2 \simeq t\left(\sigma_j Z_j - \sum_{l=1}^{k}\gamma_l(0)\sigma_l Z_l\right)^2$ which is linear in $\sigma^2 t$, to leading order and which has an expectation of

$$E\left[\|\mathbf{L}(t)-\mathbf{L}(0)\|^2\right] \simeq t\sum_{j=1}^{k} E\left[\left(\sigma_j Z_j - \sum\nolimits_{l=1}^{k}\gamma_l(0)\sigma_l Z_l\right)^2\right]$$

$$= t\sum_{j=1}^{k}\sigma_j^2\left[1-2\gamma_j(0)+k\gamma_j^2(0)\right]. \tag{9}$$

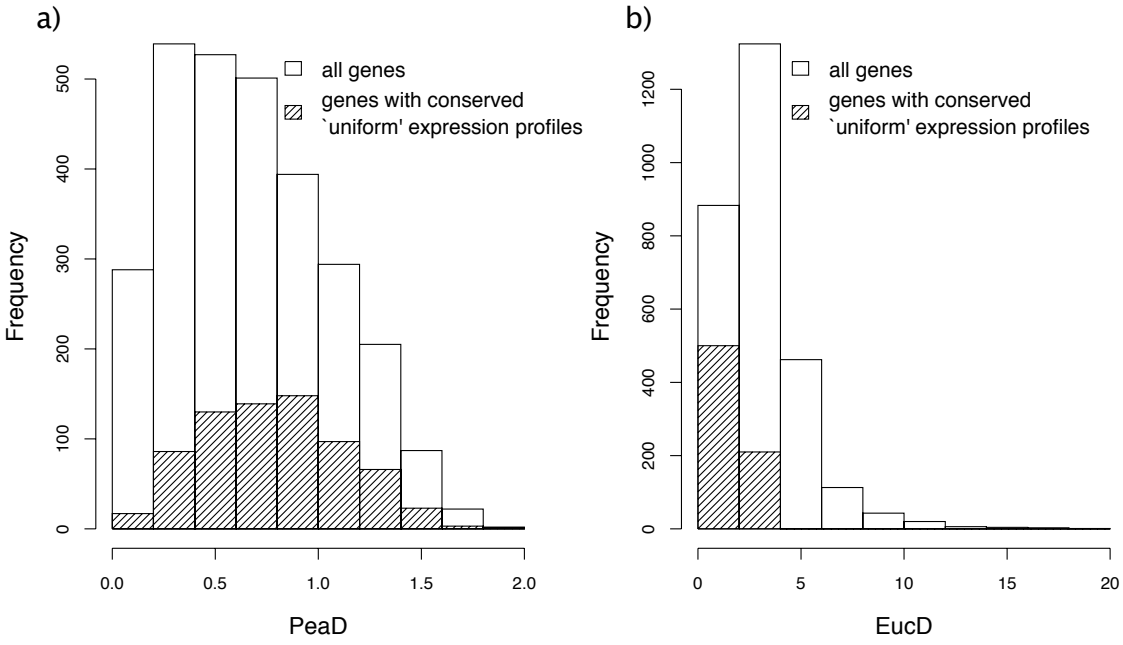

FIGURE S1.—The distribution of expression divergence values for those genes with a uniform pattern expression that this is conserved across species, versus the distribution for all genes for (a) Pearson and (b) Euclidean distances for human-mouse. Only the results from MAS5 normalization are shown; qualitatively similar results were obtained with RMA.

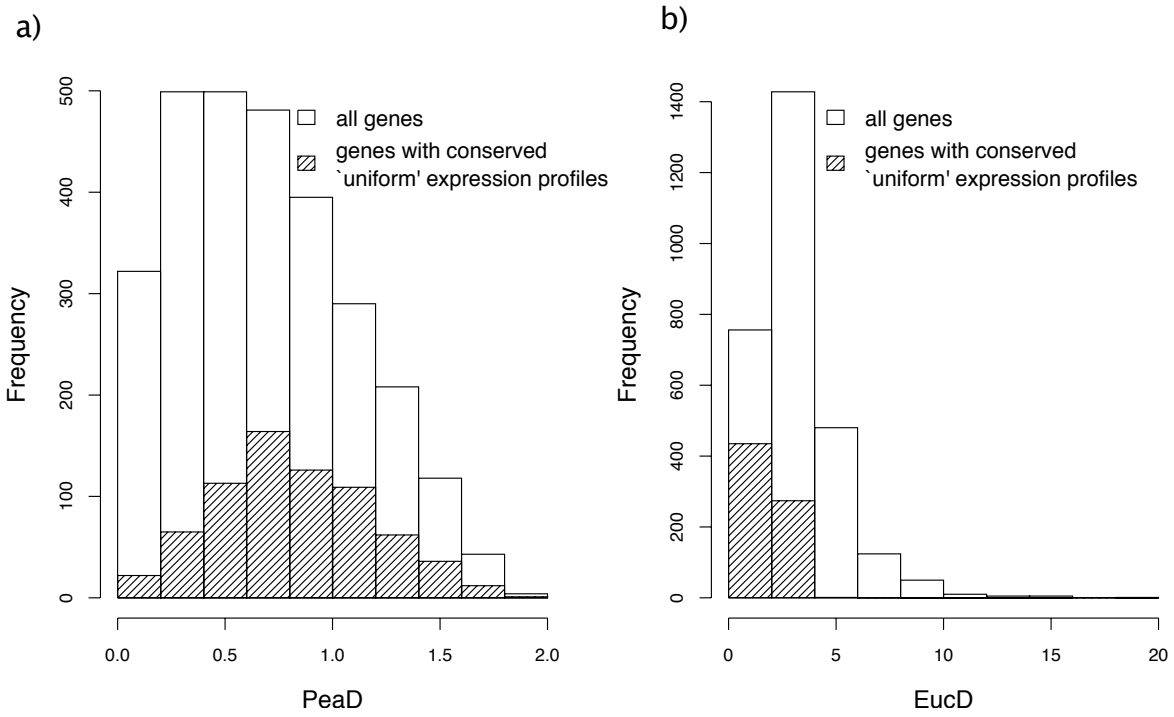

FIGURE S2.—The distribution of expression divergence values for those genes with a uniform pattern expression that this is conserved across species, versus the distribution for all genes for (a) Pearson and (b) Euclidean distances for human-rat. Only the results from MAS5 normalization are shown; qualitatively similar results were obtained with RMA.