DOI: 10.1534/genetics.104.037135

Toward a Neutral Evolutionary Model of Gene Expression

Philipp Khaitovich,* Svante Pääbo* and Gunter Weiss*,†,1

**Max Planck Institute for Evolutionary Anthropology, D-04103 Leipzig, Germany and †WE Informatik, Bioinformatik, University of Düsseldorf, D-40225 Düsseldorf, Germany*

Manuscript received October 6, 2004
Accepted for publication March 7, 2005

ABSTRACT

We introduce a stochastic model that describes neutral changes of gene expression over evolutionary time as a compound Poisson process where evolutionary events cause changes of expression level according to a given probability distribution. The model produces simple estimators for model parameters and allows discrimination between symmetric and asymmetric distributions of evolutionary expression changes along an evolutionary lineage. Furthermore, we introduce two measures, the skewness of expression difference distributions and relative difference of evolutionary branch lengths, which are used to quantify deviation from clock-like behavior of gene expression distances. Model-based analyses of gene expression profiles in primate liver and brain samples yield the following results: (1) The majority of gene expression changes are consistent with a neutral model of evolution; (2) along evolutionary lineages, upward changes in expression are less frequent but of greater average magnitude than downward changes; and (3) the skewness measure and the relative branch length difference confirm that an acceleration of gene expression evolution occurred on the human lineage in brain but not in liver. We discuss the latter result with respect to a neutral model of transcriptome evolution and show that a small number of genes expressed in brain can account for the observed data.

THE neutral theory of molecular evolution states "that at the molecular level evolutionary changes and polymorphisms are mainly due to mutations that are nearly enough neutral with respect to natural selection that their behavior and fate are mainly determined by mutation and random drift" (Kimura 1983, p. 34). While assumptions and details are still under discussion, the neutral theory has proven immensely fruitful in that it provides a null model for evolutionary analyses of molecular genetic data (for an overview see, *e.g.*, Swofford *et al.* 1996; Durrett 2002; Balding *et al.* 2003).

Gene expression has been studied within as well as between species in various organisms (Jin *et al.* 2001; Enard *et al.* 2002; Oleksiak *et al.* 2002; Su *et al.* 2002; Caceres *et al.* 2003; Cheung *et al.* 2003; Rifkin *et al.* 2003; Schadt *et al.* 2003) and some authors have discussed what fraction of genes may evolve under neutrality (Hsieh *et al.* 2003; Rifkin *et al.* 2003). Recently, we studied the expression evolution in tissues from primates and mice (Khaitovich *et al.* 2004). We found that transcriptome divergence between species correlates positively with intraspecies expression diversity and accumulates approximately linearly with time. We also found that the rates of transcriptome divergence between a set of expressed pseudogenes and intact genes do not differ significantly. These observations led us to suggest that a neutral model of evolution may apply to the transcriptome, *i.e.*, that the majority of genes expressed in a certain tissue change over evolutionary time as the result of stochastic processes that are limited in their extent by negative selection rather then as the result of positive Darwinian selection.

By considering gene expression as a quantitative character, others (Rifkin *et al.* 2003; Gu 2004) have used Brownian motion models (Edwards and Cavalli-Sforza 1964; Felsenstein 1973; Lande 1976; Lynch and Hill 1986) to describe the evolution of gene expression. Here, we introduce a stochastic model that describes neutral changes of gene expression over evolutionary time as a compound Poisson process. Although this model considers only mutations in *cis*-regulatory elements explicitly and ignores interactions between genes, it enables us to describe the evolutionary process in more detail and explain and quantify some general phenomena of transcriptome evolution. We illustrate the use of the model by analyzing gene expression profiles from primate liver and brain.

MODELING THE EVOLUTION OF GENE EXPRESSION

General model: We propose a stochastic model of gene expression evolution that describes mutations on the DNA level as a Poisson process and the effect of a mutation on a gene's expression by some probability distribution. The mutations (and their effects) are assumed mutually independent. The effects of mutations

[1] *Corresponding author:* WE Informatik, Bioinformatik, University of Düsseldorf, Universitätsstrasse 1, D-40225 Düsseldorf, Germany. E-mail: weiss@cs.uni-duesseldorf.de

are thought to act multiplicatively on expression intensities; *i.e.*, the expression level after a mutation occurred is a multiple of the level before mutation of the sequence. This means that the relative amount of change in expression caused by a mutation is independent from the absolute expression level. To make the model additive, we replace expression levels by their logarithm. Finally, we assume that over evolutionary time there is no bias for increasing or decreasing a gene's expression level, *i.e.*, that evolution is not directional. In mathematical terms the evolutionary process of gene expression is a compound Poisson process with independent increments. More formally, let $M(t)$ be a random variable describing the number of mutations occurring in the regulatory region of some gene in some time interval of length t. Here, we consider time on an evolutionary scale, *i.e.*, real time scaled by the mutation rate such that time corresponds to branch lengths in an evolutionary tree. Then, the expression value on the log scale $Y(t)$ after t units of scaled time is given as

$$Y(t) = Y(0) + \sum_{i=1}^{M(t)} X_i,$$

where X_i denotes the effect of mutation i on log expression value. The random variables X_i are independent and follow some distribution with zero mean [$\mu(X) = 0$], which we specify later. Thus, $Y(t)$ defines a compound Poisson process. Since we are concerned mainly with comparative data, we describe differences in expression between two samples before analyzing the model in more detail. These expression differences between two samples are the data usually measured using either oligonucleotide or cDNA arrays. Let $Z_{1,2}$ describe a gene's expression difference between two samples that have evolved independently on branches of length t_1 and t_2 from a common ancestor. Then,

$$Z_{1,2} = Y_1(t_1) - Y_2(t_2) = \sum_{i=1}^{M(t_1)} X_i - \sum_{j=1}^{M(t_2)} X_j,$$

since the common ancestry guarantees that $Y_1(0) = Y_2(0)$. Note that we consider evolutionary time such that t_1 and t_2 may differ. The random variable $Z_{1,2}$ defines the difference of two independent compound Poisson processes now subject to further investigation. A closed formula for the density function of $Z_{1,2}$ does not exist. However, moments can be derived using characteristic functions defined by $\phi_X(\theta) = E(e^{i\theta X})$. The characteristic function of $Z_{1,2}$ is given by $\phi_{Z_{1,2}}(\theta) = \phi_{Y_1}(\theta) \cdot \overline{\phi_{Y_2}(\theta)}$, where $\phi_{Y_j}(\theta) = \phi_{N(t_j)}(\phi_X(\theta)) = \exp(t_j(\phi_X(\theta) - 1))$ and $\overline{\phi_\bullet(\theta)}$ is the complex conjugate of $\phi_\bullet(\theta)$ (for details see, *e.g.*, Feller 1957).

Let $\mu(X)$ denote the mean and $\mu_k(X)$ the kth (central) moment of random variable X and define its coefficients of skewness and kurtosis as $\gamma_1(X) = \mu_3(X)/(\mu_2(X))^{3/2}$ and $\gamma_2(X) = \mu_4(X)/(\mu_2(X))^2$, respectively. Computation of the corresponding quantities for $Z_{1,2}$ is straightforward using $\mu_k(Z_{1,2}) = i^k \phi^{(k)}_{Z_{1,2}}(\theta)|_{\theta=0}$, where $\phi^{(k)}_{Z_{1,2}}(\theta)$ denotes the kth derivative of $\phi_{Z_{1,2}}(\theta)$. Moments of $Z_{1,2}$ can be expressed in terms of characteristics of the distribution of X:

$$\mu(Z_{1,2}) = \mu(X)(t_1 - t_2) = 0,$$

$$\mu_2(Z_{1,2}) = \mu_2(X)(t_1 + t_2), \tag{1}$$

$$\gamma_1(Z_{1,2}) = \gamma_1(X)\frac{t_1 - t_2}{(t_1 + t_2)^{3/2}}, \tag{2}$$

$$\gamma_2(Z_{1,2}) = 3 + \frac{\gamma_2(X)}{t_1 + t_2}. \tag{3}$$

The mean of $Z_{1,2}$ equals zero, since we assumed a zero mean distribution for X. The variance of $Z_{1,2}$ grows linearly with the sum of branch lengths and the coefficient of skewness of $Z_{1,2}$ depends on a scaled difference of branch lengths. We note that for $t_1 + t_2$ large, the moment ratios of $Z_{1,2}$ converge to those of a normal distribution; *i.e.*, the limiting case of this model is a Brownian motion.

Specifying the effect of a random mutation: Up to here, we considered a general distribution of X. Below we study two special cases of mutational effect distributions, namely normally distributed effects and effects following an extreme value distribution. The normal distribution corresponds to the symmetric case where a random mutation causes equally likely a decrease and an increase in expression. Since we assume that the mean is zero [$\mu(X) = 0$], this distribution is uniquely specified by its variance, $\mu_2(X) = \sigma^2$ [$\gamma_1(X) = 0, \gamma_2(X) = 3$]. An extreme value distribution with parameters α and β (Johnson *et al.* 1995) is used to describe a situation where a mutation is more likely to reduce the expression of the gene (see Figure 1). Here, expression evolves with more frequent but smaller downward jumps compensated by fewer upward jumps of bigger size. Moments and moment ratios of this distribution are: $\mu(X) = \alpha + \beta\lambda$, $\mu_2(X) = \pi^2\beta^2/6$, $\gamma_1(X) = 12\sqrt{6}\zeta(3)/\pi^3 \approx 1.13955\ldots$, $\gamma_2(X) = 27/5$, where $\zeta(\cdot)$ is the ζ-function and $\lambda \approx 0.57721\ldots$ is the Euler-Mascheroni constant. We set $\alpha = -\beta\lambda$ to assure a zero mean for X. Thus, this extreme value distribution is specified by a single parameter β. Another possibility would be to use a negatively skewed distribution, *e.g.*, a mirrored version of an extreme value distribution. However, as we show later such a model is not consistent with the data and we do not pursue that case further.

Estimating parameters: Equations 1–3 yield estimators for the model parameters via the method of moments. The length of the evolutionary path between the two samples is estimated via Equation 3:

$$t_1 + t_2 = \frac{\gamma_2(X)}{\gamma_2(Z_{1,2}) - 3}.$$

An estimator for the variance of the effect distribution X is derived from Equations 1 and 3:

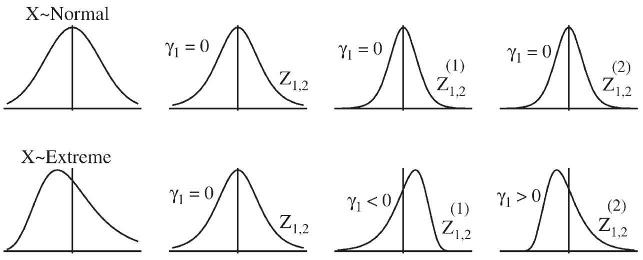

Figure 1.—Predicted distributions of expression differences between samples 1 and 2 for two different mutational effect distributions X, given $t_1 = t_2$. The top row indicates predictions for a normally distributed effect model: Distributions of expression differences of all genes ($Z_{1,2}$), of sample 1-intermediate genes only ($Z_{1,2}^{(1)}$), and of sample 2-intermediate genes only ($Z_{1,2}^{(2)}$) are all symmetric ($\gamma_1 = 0$). The bottom row shows predictions for a positively skewed extreme value distribution effect model: Distribution of differences over all genes is symmetric; distributions of sample 1-intermediate and sample 2-intermediate genes are negatively ($\gamma_1 < 0$) and positively ($\gamma_1 > 0$) skewed, respectively, and are therefore indicative for an asymmetric effect model distribution.

$$\mu_2(X) = \frac{\mu_2(Z_{1,2})(\gamma_2(Z_{1,2}) - 3)}{\gamma_2(X)}.$$

Asymmetric cases with nonzero skewness of X ($\gamma_1(X) \neq 0$), such as the extreme value distribution, permit us to estimate the branch lengths separately using Equations 1–3:

$$t_{1/2} = \frac{1}{2}\frac{\gamma_2(X)}{\gamma_2(Z_{1,2}) - 3}\left(1 \pm \frac{\gamma_1(Z_{1,2})}{\gamma_1(X)}\sqrt{\frac{\gamma_2(X)}{\gamma_2(Z_{1,2}) - 3}}\right). \qquad (4)$$

Now assume that additionally to the two samples we have a third sample that serves as an outgroup (see Figure 2). Let $\mu_2(Z_{i,j})$ denote the variance of the difference $Z_{i,j}$ between samples i and j. We can use the outgroup data to construct estimators for branch lengths for the normal and the extreme value distribution case. Since $\mu_2(Z_{j,3}) = \mu_2(X)(t_j + t_0 + t_3)$ for $j = 1, 2$ (see Figure 2), it is easy to verify that

$$t_1 = \frac{1}{2}(\mu_2(Z_{1,2}) + \mu_2(Z_{1,3}) - \mu_2(Z_{2,3}))(\mu_2(X))^{-1},$$

$$t_2 = \frac{1}{2}(\mu_2(Z_{1,2}) - \mu_2(Z_{1,3}) + \mu_2(Z_{2,3}))(\mu_2(X))^{-1}.$$

Note that for $t_1 + t_2$ large all the above estimators do not behave well, since the coefficient of skewness $\gamma_1(Z_{1,2})$ and the kurtosis excess $\gamma_2(Z_{1,2}) - 3$ both converge to zero. In this situation one is left with $\mu_2(Z_{1,2})$ as a measure of evolutionary distance, which is scaled by the (unknown) variance of the mutational effect $\mu_2(X)$.

Another quantity of interest is the relative difference in branch lengths, $\tau_{1,2}$ (as defined below), which provides information about the clock-likeness of evolutionary trees built from expression differences. This measure is independent of any choice of the mutational effect distribution X. Its estimator involves only second moments and thus is comparably robust:

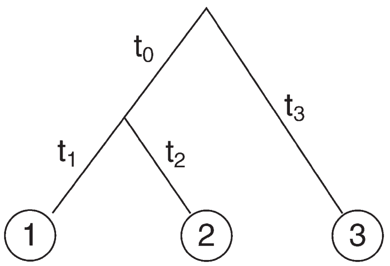

Figure 2.—Evolutionary tree of three taxa.

$$\tau_{1,2} = \frac{t_1 - t_2}{t_1 + t_2} = \frac{\mu_2(Z_{1,3}) - \mu_2(Z_{2,3})}{\mu_2(Z_{1,2})}. \qquad (5)$$

To estimate the moments of $Z_{i,j}$ from current data we assume that an expression profile with N genes measured represents a set of N independent realizations of the described evolutionary process. This assumption neglects any *trans*-effects on gene expression as well as any interactions of genes, both of which surely exist.

To judge the performance of the proposed estimators we generated data along a three-species tree (see Figure 2) under our model for several combinations of parameters via computer simulation and applied the estimation procedure to the artificial data. More precisely, we used the following parameter settings (see Figure 3): $t_1 = t_2 = 1$ (cases a, b, and c) and $t_1 = 1.5$, $t_2 = 0.5$ (cases d, e, and f); $\mu_2(X) = 0.25$ (cases a and d), $\mu_2(X) = 0.33$ (cases b and e), $\mu_2(X) = 0.5$ (cases c and f); $t_0 = 1.0$, $t_3 = 2.0$ (all cases); and data generated for $N = 2000$, 5000, and 10,000 genes (indexed by 2, 5, and 10, respectively). The results of this analysis are summarized in Figure 3, where we assumed an extreme value distribution for the mutational effect. Indicated are mean and 95% probability intervals for the estimates. We observe a bias in the parameter estimates that decreases with the number of genes. The estimate of $\mu_2(X)$ has a small relative bias of 1% ($N = 2000$), 0.5% ($N = 5000$), and 0.3% ($N = 10{,}000$). The estimates of the times t_1 and t_2 are biased upward (relative bias of 6, 3, 1.5%, respectively). As expected, we find the range of estimates to decrease substantially with the number of genes analyzed. The results of the very same analysis under a normally distributed mutation effect are completely comparable (data not shown).

Discriminating between symmetric and asymmetric effect models: Given that the lengths of the evolutionary branches leading to the two samples are different ($t_1 \neq t_2$), the skewness $\gamma_1(Z_{1,2})$ can be used to discriminate between symmetric and asymmetric effect models for the mutational effect X, since this quantity is expected to differ from zero only if an asymmetric mutational model applies (see Equation 2). In contrast, if we know that branch lengths are the same ($t_1 = t_2$), the coefficient

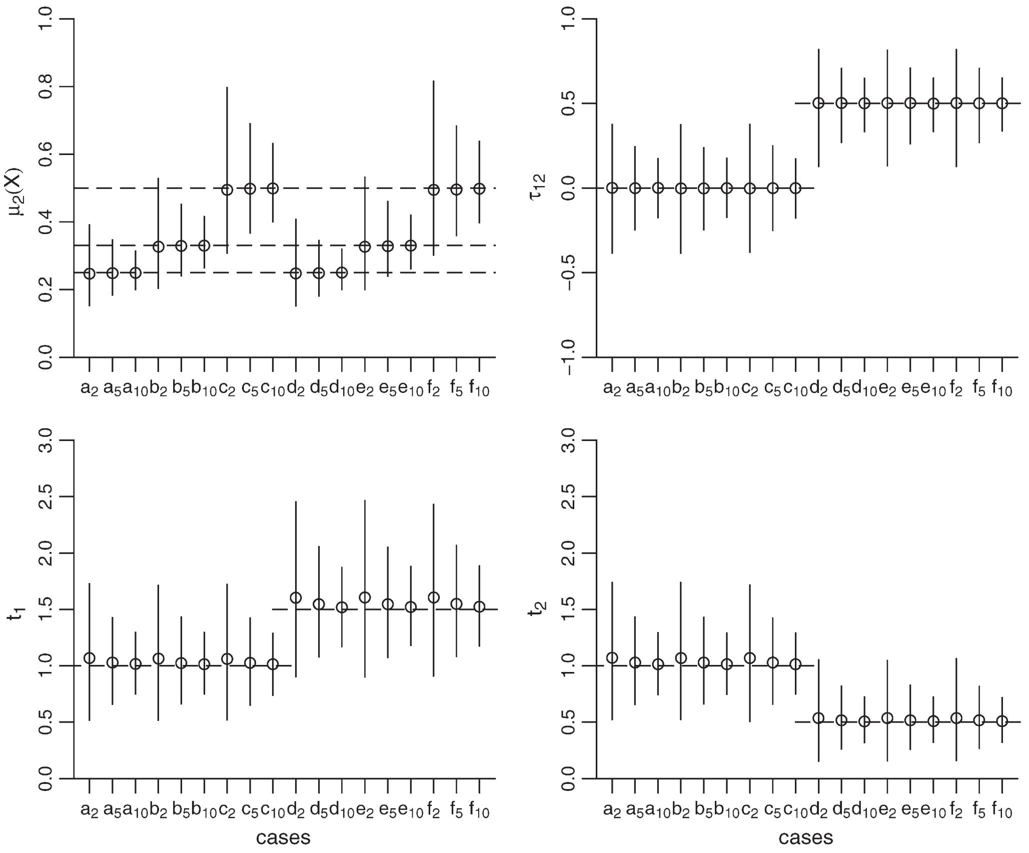

Figure 3.—Performance of estimators for indicated parameters. Data were generated under the proposed model with an extreme value mutation effect distribution along a three-species tree (see Figure 2), using six parameter sets (cases a–e). The number of genes (in thousands) simulated per data set is given by the index of the cases. Dashed horizontal lines correspond to parameters used for simulation. Means (circles) and 95% probability intervals (solid vertical lines) generated for the estimators from 10,000 simulated data sets per case are shown.

of skewness will be zero [$\gamma_1(Z_{1,2}) = 0$] independently of the model for X.

We construct quantities that discriminate between the two effect models independently from assumptions about evolutionary branch lengths and especially for the case $t_1 = t_2$. If we know the ancestral state of a gene's expression, this will be straightforward since we can observe changes in expression and their direction more directly. An indirect way to incorporate information about the direction of the expression changes is the use of data from an additional sample, sample 3 say, that is known to be an outgroup to samples 1 and 2 (see Figure 2). The outgroup is used to classify genes into three categories defined by the order of expression values. We call a gene "sample j-intermediate" ($j = 1, 2, 3$) if its expression value in sample j lies in between the expression levels of the remaining two samples. Of specific interest to the problem of discrimination among effect models are the sample 1- and sample 2-intermediate gene classes. The class of sample 1-intermediate genes is enriched with genes where changes on the evolutionary lineage leading to sample 2 predominantly caused the difference in expression between samples 1 and 2, while, for sample 2-intermediate genes, this difference is mainly generated by changes in sample 1. This is the case, because the intermediate expression value of a sample is more likely to be close to the ancestral expression state of samples 1 and 2 as it is bounded by the expression values of the two other samples. We study the distribution of the difference between expression values of samples 1 and 2 for sample j-intermediate genes $Z^{(j)}_{1,2}$, assuming $t_1 = t_2$. Unfortunately, we are not aware of analytical results for these distributions. Therefore, simulations in an even wider parameter range than described in the previous section were invoked to verify the following characteristics of the distributions. Given a symmetric distribution for X, the symmetry of the involved Poisson processes carries over to all three distributions, $Z^{(j)}_{1,2}$. However, this picture changes if the mutational effect distribution X is asymmetric. Given $t_1 = t_2$, the distribution of "outgroup-intermediate" genes $Z^{(3)}_{1,2}$ is still symmetric, while the distribution of sample 1-intermediate genes has a negative coefficient of skewness and that of the sample 2-intermediate genes follows a positively skewed distribution. Thus, we can use measures of skewness for distributions of expression differences for classified genes to discriminate between models. Figure 1 shows qualitatively expected distributions of expression differences for all genes ($Z_{1,2}$) as well as for sample 1- and sample 2-intermediate genes [$Z^{(1)}_{1,2}$ and $Z^{(2)}_{1,2}$, respectively] for both effect models for X.

ANALYSIS OF PRIMATE EXPRESSION PROFILES

Preprocessing of the microarray data: We analyzed four gene expression data sets from several primate

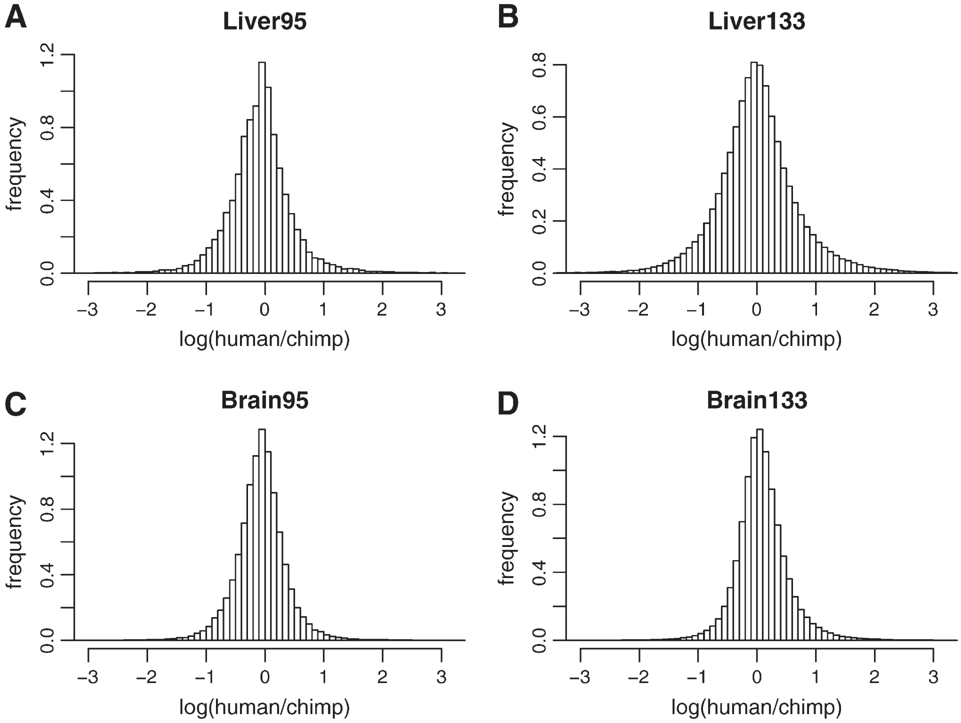

FIGURE 4.—Distribution of human-chimpanzee expression differences (on log scale) from liver95 (A), liver133 (B), brain95 (C), and brain133 (D).

species using the proposed model of neutral evolution on the transcriptome level. The first two data sets consist of liver and brain data, correspondingly collected using Affymetrix HG U95Av2 arrays. The liver data set was collected from three humans, three chimpanzees, and one orangutan with two measurements for each individual (ENARD *et al.* 2002). The brain data set consists of expression profiles from six humans, three chimpanzees, and one orangutan (KHAITOVICH *et al.* 2004). We refer to these data sets as liver95 and brain95, respectively. The third and fourth data sets comprise expression profiles from six human and five chimpanzee samples in brain and liver, respectively, and one orangutan brain sample, but five orangutan liver samples (KHAITOVICH *et al.* 2005). These data were collected with Affymetrix *U133plus2* arrays and are denoted by liver133 and brain133. To minimize artifacts that result from hybridizing chimpanzee samples to human arrays, we masked all oligonucleotide probes where DNA sequence did not match perfectly between the chimpanzee and the human genome as described elsewhere (KHAITOVICH *et al.* 2004). Each data set was normalized and gene expression intensity values were calculated using the Bioconductor rma function (IHAKA and GENTLEMAN 1996; BOLSTAD *et al.* 2003). Note that we consider log-transformed intensities as our data. Finally, within each data set, we restricted our analysis to probe sets expressed significantly above background in all samples as gauged by default detection *P*-value (Affymetrix Microarray Suite v5.0). In total, the analyzed data consist of measurements from 1971 probe sets (liver95), 1998 probe sets (brain95), 8005 probe sets (liver133), and 10,414 probe sets (brain133). Figure 4 illustrates the data in terms of their human-chimpanzee difference distributions computed over all individual pairs and probe sets on log scale.

Squared expression differences accumulate approximately linearly with time: If the majority of genes evolves neutrally with respect to their expression (KHAITOVICH *et al.* 2004), our model predicts a linear relationship between time since divergence of the species and the variance of expression differences. To estimate transcriptome divergence between two species, we computed the variance of expression differences for each sample pair from the two species and averaged over pairs. Confidence regions were constructed by bootstrapping over individuals and genes 10,000 times, taking 2.5 and 97.5% quantiles of the bootstrap distribution as limits. Within species, transcriptome diversity was estimated by the averaged variance for within-species comparisons of humans or chimpanzees. Bootstrapping 10,000 times over genes assessed uncertainty in these estimates (Table 1). Averages of pairwise gene expression variances with corresponding confidence intervals plotted against estimates of divergence times based on DNA sequence data (GLAZKO and NEI 2003) are shown in Figure 5. An approximately linear relationship holds for all four data sets even though the variation of the estimates is considerable. All data sets have nonzero intercepts of the regression lines with the *y*-axis. The apparent excess of gene expression variance over that expected from DNA sequence data may be due to expression differences caused by nongenetic factors such as experimental variation and environmental effects. This effect is of similar

TABLE 1

Estimates of variances with corresponding 95% bootstrap intervals (numbers in parentheses) from expression differences for four data sets

Comparison	Liver95	Brain95	Liver133	Brain133
Within humans	0.103 (0.094; 0.114)	0.119 (0.112; 0.127)	0.254 (0.168; 0.344)	0.093 (0.055; 0.133)
Within chimpanzees	0.147 (0.133; 0.162)	0.094 (0.085; 0.105)	0.417 (0.256; 0.579)	0.103 (0.072; 0.141)
Within orangutans	—	—	0.310 (0.188; 0.429)	—
Human *vs.* chimpanzee	0.284 (0.245; 0.330)	0.163 (0.130; 0.206)	0.503 (0.412; 0.608)	0.194 (0.171; 0.218)
Human *vs.* orangutan	0.458 (0.403; 0.527)	0.380 (0.315; 0.445)	0.713 (0.600; 0.843)	0.465 (0.426; 0.501)
Chimpanzee *vs.* orangutan	0.450 (0.402; 0.501)	0.312 (0.258; 0.374)	0.717 (0.611; 0.832)	0.412 (0.375; 0.451)

The first column indicates the individual samples compared.

magnitude in all data sets with the exception of liver133, where it is larger. The overall divergence on the transcriptome level appears larger in liver than in brain. The slopes of the regressions are very similar, only the regression slope of brain95 is about one-third smaller.

Estimating model parameters: The separate estimation of the length of evolutionary branches (mutation rate times real time) and of the variance of the mutational effect distribution depends on estimation of moment ratios with third and fourth power terms. The coefficients of skewness and kurtosis were estimated as averages over the corresponding estimates from appropriate pairwise comparisons (Table 2). The uncertainty attached to these estimates was judged by bootstrapping 10,000 times over individuals and genes. The distribution of differences between humans and chimpanzees has a positive skew in all four data sets. For the two liver data sets and brain95 the 95% confidence intervals include zero. This is not the case for brain133, where the lower limit of the 95% confidence interval is well above zero. According to our model, these observations can be translated into length (and rate) differences of the human and the chimpanzee evolutionary branches from the most recent common ancestor (see Equation 2). Additionally, it excludes a symmetric distribution for the mutational effects in our model. We translate empirical moments of the human-chimpanzee comparison in parameter estimates for both mutational effect distributions. Table 3 shows these estimates together with their 95% confidence intervals based on 10,000 bootstraps. The estimates of branch lengths and mutational effect variances derived from different data sets are in good agreement, but also reflect the different amounts of within-species variation. The relatively larger estimates for liver133 are the result of its unusual large within-species variation (and regression intercept, Figure 5). Under our model, assuming that the extreme value distribution applies for the mutational effect, the positive

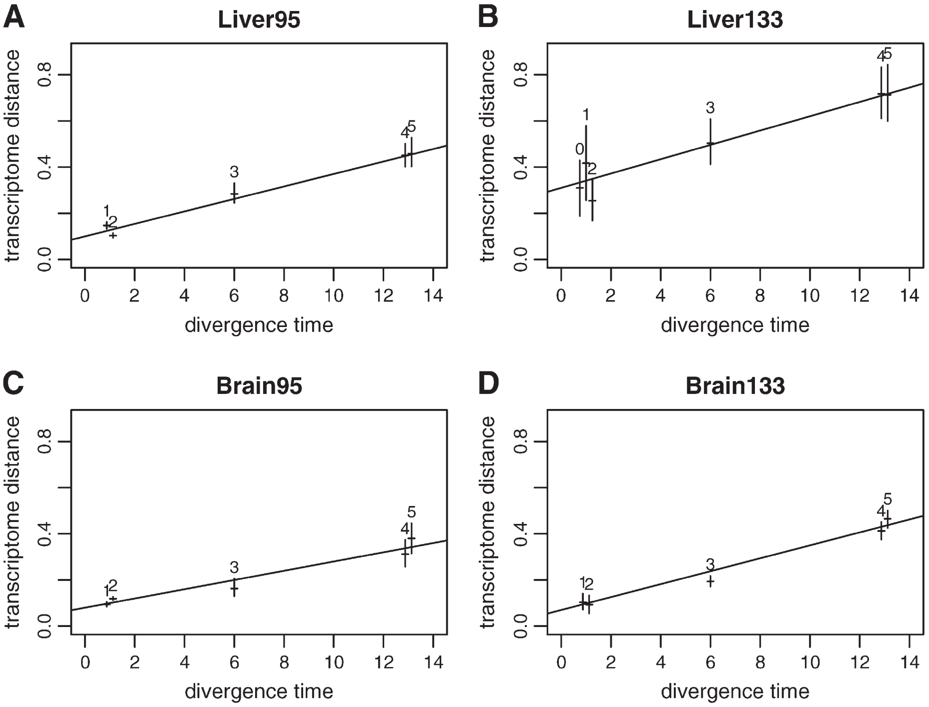

Figure 5.—Transcriptome distance measured as averaged pairwise variance of expression differences (*y*-axis) as a function of time since divergence in millions of years (*x*-axis) (Glazko and Nei 2003), for (A) liver95, (B) liver133, (C) brain95, and (D) brain133. Numeral code: 0, comparison within orangutan; 1, within chimpanzees; 2, within humans; 3, between human and chimpanzee; 4, between chimpanzee and orangutan; 5, between human and orangutan.

TABLE 2

Estimates of skewness $\gamma_1(Z_{H,C})$ and kurtosis $\gamma_2(Z_{H,C})$ with corresponding 95% bootstrap intervals (numbers in parentheses) for human-chimpanzee expression differences for four data sets

	Liver95	Brain95	Liver133	Brain133
Skewness	0.35 (−0.08; 0.81)	0.44 (−0.19; 0.96)	0.21 (−0.04; 0.43)	0.59 (0.33; 0.83)
Kurtosis	7.46 (5.97; 9.35)	7.91 (5.92; 10.02)	6.63 (5.70; 7.70)	8.10 (7.20; 9.06)

skew of the distributions of *Z* for all data transforms into longer evolutionary branches to the human than to the chimpanzee. While the ratio is ~1.8 (and not significantly different from 1) for liver95, brain95, and liver133 data sets, the human branch is significantly longer than the chimpanzee branch for the brain133 data set [ratio 3.32, bootstrap interval (1.88; 6.64)].

A test for differences in evolutionary branch lengths: The relative difference of evolutionary branch lengths $\tau_{1,2}$ carries information about the clock-likeness of evolutionary trees. A neutral model of expression evolution would predict trees that are approximately clock-like. By scaling the difference of branch lengths by their sum, the statistic gets independent of the specific choice of mutational effect distribution *X*. We estimate the relative length differences of the human branch to the chimpanzee branch using orangutan as an outgroup. To this end, within each data set, we computed $\tau_{H,C}$ for all possible trios of a human, a chimpanzee, and an orangutan with variances as appropriate distance measures between taxa (see Equation 5). As estimates we report the averages over these values within each data set. Uncertainty in the estimates is reported as 95% confidence intervals based on 10,000 bootstraps. The estimated relative length differences of human branch to chimpanzee branch are very close to zero in the two liver data, but substantially positive in the two brain data sets. The 95% bootstrap interval of $\tau_{H,C}$ for brain133 marginally excludes zero (Table 4). Thus, we find evidence that the length of the human branch from the most recent common ancestor is longer than the chimpanzee branch in the brain133 data set both using the orangutan as an outgroup and using the skewness of the distribution of human-chimpanzee expression differences. Therefore, clock-like evolutionary trees are applicable to describe human-chimpanzee expression differences in both liver data sets and to a lesser extent in the brain95 data set, but not in the brain133 data set.

Effects of mutations are not symmetric: The observation that the distribution of human-chimpanzee differences is significantly skewed in the brain133 data set rules out any symmetric mutational effect distribution like, for instance, a normal distribution. It also calls into question the appropriateness of Brownian motion models for description of transcriptome evolution. In contrast, our model using the extreme value distribution of the mutational effect describes the skewed distribution of human-chimpanzee expression differences observed in the brain133 data set well. In addition, the extreme value distribution effect model predicts that distribution of human-chimpanzee expression differences is skewed with a negative skew for the human-intermediate genes and positively skewed for the chimpanzee-intermediate genes (Figure 1). To test this prediction, we investigated the shape of the difference distributions after genes are grouped into the human-intermediate and chimp-intermediate class. Again the orangutan serves as the outgroup used for classification. We computed skewness for both gene classes using all possible pairs of a human and a chimpanzee sample within each data set and report the average of these estimates (Table 5). We used Pearson's coefficient of skewness (a robust

TABLE 3

Estimates of model parameters with corresponding 95% bootstrap intervals (numbers in parentheses) from human-chimpanzee expression differences for four data sets

Parameter	Liver95	Brain95	Liver133	Brain133
		Normal distribution		
$\mu_2(X) = \sigma^2$	0.42 (0.26; 0.66)	0.25 (0.15; 0.37)	0.58 (0.45; 0.74)	0.33 (0.26; 0.40)
$t_H + t_C$	0.71 (0.50; 1.03)	0.75 (0.49; 1.34)	0.92 (0.69; 1.25)	0.60 (0.50; 0.74)
		Extreme value distribution		
$\mu_2(X) = \pi^2\beta^2/6$	0.24 (0.14; 0.35)	0.14 (0.08; 0.20)	0.32 (0.25; 0.42)	0.18 (0.15; 0.22)
$t_H + t_C$	1.28 (0.90; 1.84)	1.34 (0.88; 2.41)	1.66 (1.24; 2.24)	1.08 (0.91; 1.32)
t_H	0.83 (0.59; 1.17)	0.87 (0.51; 1.44)	1.07 (0.67; 1.66)	0.83 (0.66; 1.06)
t_C	0.45 (0.18; 0.91)	0.47 (0.07; 1.37)	0.58 (0.40; 0.81)	0.25 (0.14; 0.39)

TABLE 4

Relative differences of human to chimpanzee branch lengths ($\tau_{H,C}$) with corresponding 95% bootstrap intervals (numbers in parentheses) for four data sets

	Liver95	Brain95	Liver133	Brain133
$\tau_{H,C}$	0.03 (−0.15; 0.23)	0.36 (−0.10; 0.76)	−0.009 (−0.23; 0.20)	0.28[a] (0.02; 0.52)

[a] Estimates with 95% bootstrap intervals excluding zero.

measure based on the scaled difference between mean and median) because the numbers of genes in both classes are in the range of a few hundred for the liver95 and brain95 data sets and differ substantially between sample pairs. All four data sets yield human-intermediate distributions with negative skew and chimp-intermediate distributions skewed in the opposite direction. This result is not expected if a symmetric distribution of the mutational effect applies, *i.e.*, if up- and downregulations of genes are equally frequent and of equal average magnitude. However, if the distribution of the mutational effect follows a positively skewed distribution like the extreme value distribution, the observed pattern of the human- and chimpanzee-intermediate distributions matches the expectation. Three out of eight 95% confidence intervals constructed around the estimates do not include zero. Thus, we conclude that an evolutionary model with a positively skewed mutational effect distribution X is superior to the models based on symmetric effect distributions in explaining the data.

DISCUSSION AND CONCLUSION

We introduce a stochastic model that describes gene expression evolution where the observable difference in expression of a gene between two samples is generated by the difference of two independent compound Poisson processes. A method of moments approach yields simple estimators for the model parameters involving second, third, and fourth moments of the distribution. We assume that genes evolve independently of each other. While this is reasonable when only *cis*-effects are considered, this assumption gets more problematic when we consider the whole transcriptome where gene products may affect target genes via *trans*-effects. As long as *trans*-effects are restricted to single genes our approach will be valid. Only when many genes are affected in the same way by one and the same gene will our parameter estimates be biased. However, recent studies in flies (Wittkopp *et al.* 2004) and humans (Morley *et al.* 2004) indicate that evolution of *cis*- and single-gene *trans*-effects are predominant. In future work we hope to include all kinds of *trans*-effects.

Despite the simplicity of the proposed model, it appears useful in several respects. For example, the second moment (variance) can be used as an additive distance measure for evolutionary branches to construct phylogenetic trees from expression data, and the relative difference of branch lengths ($\tau_{1,2}$) measures evolutionary acceleration on a specific lineage. The advantage of $\tau_{1,2}$ is that it is independent of the choice of mutational effect distribution. The disadvantage is that it relies on the analysis of an outgroup species. This might be problematic since suitable outgroup species may not exist or be unobtainable, or their genome sequences may not be determined so that hybridization artifacts cannot be fully controlled for (see *Preprocessing of the microarray data*). Potentially, the usage of cDNA arrays or the like (Rise *et al.* 2004) and the availability of customized oligonucleotide arrays will widen the applicability of our approach. The use of the third moment (skewness) of expression differences provides a way to directly detect evolutionary accelerations without the recourse to an outgroup since the skewness of the expression difference distribution quantifies branch length differences (see Equation 3). This is possible when the distributions of mutational effects are themselves skewed such that they contain more downregulations than upregulations, where the former are of smaller average amplitude than the latter.

Previously, a brain-specific acceleration on the human lineage (Enard *et al.* 2002; Caceres *et al.* 2003; Gu and Gu 2003) was reported using the orangutan as an outgroup. When we apply the model to four sets of expression data from brains and livers of humans and apes,

TABLE 5

Estimates of skewness, $\gamma_1(Z^{(H)}_{H,C})$ and $\gamma_1(Z^{(C)}_{H,C})$, with corresponding 95% bootstrap intervals for human-chimpanzee expression differences after grouping genes according to the relation of expression values to the orangutan

Gene class	Liver95	Brain95	Liver133	Brain133
Human-intermediate genes	−0.23[a] (−0.35; −0.08)	−0.09 (−0.25; 0.09)	−0.14 (−0.30; 0.04)	−0.07 (−0.25; 0.16)
Chimpanzee-intermediate genes	0.07 (−0.11; 0.24)	0.14 (−0.08; 0.30)	0.30[a] (0.13; 0.44)	0.51[a] (0.37; 0.54)

[a] Estimates with 95% bootstrap intervals excluding zero.

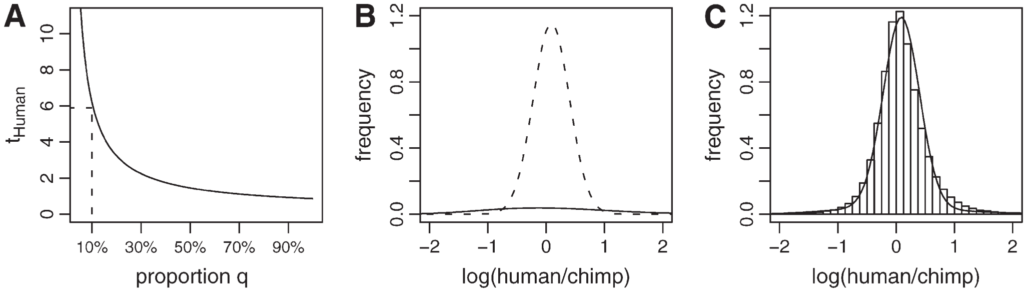

Figure 6.—Illustration of a mixture of two models to explain the significant skewness in brain133 analysis. The two models differ only in the assumed evolution. Model *A* has branch lengths t_{Human} and t_{C}; in model *B* both branches are of equal length t_{C}. (A) Solutions of the mixture of two models (see appendix, Equation A1). The branch length t_{Human} of model *A* is shown as a function of the fraction *q* of genes that evolve according to model *A*. The dashed lines indicate parameters for the example depicted in B and C. (B) Contribution of 10% of genes evolving according to model *A* (solid line) and 90% of genes evolving according to model *B* (dashed line) to the human-chimpanzee difference distribution. (C) Histogram of human-chimpanzee expression differences in brain133 and fitted mixture distribution with the same variance and skewness as brain133 data.

both the relative differences of branches $\tau_{\mathrm{H,C}}$ and the skewness $\gamma_1(Z_{\mathrm{H,C}})$ of the human-chimpanzee expression differences tend to confirm this result. Thus, while neither the liver95 nor the liver133 data set shows a significant relative difference of branch length or a skewness that is significantly different from zero, the brain133 data set shows clear evidence for an excess of gene expression changes on the human branch with both measures, and the brain95 data set, where the number of genes limits the strength of conclusion, shows a tendency toward a longer evolutionary branch leading to humans. Note, as an aside, that the observation of a significant nonzero skewness of the gene expression difference distribution in the brain133 data set raises questions about the appropriateness of Brownian motion models for the evolution of the transcriptome (Rifkin *et al.* 2003; Gu 2004) since it can not explain the finding of a skewed expression difference distribution.

The suggestion that the distribution of expression difference is significantly skewed toward more downregulations than upregulations contradicts a report by Caceres *et al.* (2003) that found an apparent excess of gene-expression upregulations in the human brain. However, since our findings indicate that downregulations are smaller in amplitude than upregulations, the cutoff criteria used by these authors are likely to restrict their analysis to the upper and lower tails of the expression difference distributions. Therefore, it is possible that the more frequent (but small) downregulations were not scored and that this caused the acceleration on the human lineage to appear to be confined to upregulations. However, further studies are needed to shed light on the molecular mechanisms underlying the observed acceleration of gene expression evolution in the human brain.

At first glance, the observation of a significant nonzero skewness in brain is also in conflict with the hypothesis postulating that the majority of evolutionary changes in gene expression are selectively neutral or nearly neutral (Khaitovich *et al.* 2004). It is also in apparent contradiction to the overall pattern of divergence in liver and brain data that suggest clock-like behavior consistent with neutrality (Figure 5). However, there may be no fundamental opposition between these observations if a small fraction of genes could account for the acceleration seen in brain. We addressed this question by estimating the fraction of human-specific gene expression changes required to yield the observed data. To do this, we restrict our attention to a simple mixture of two models: A fraction $1 - q$ of genes evolves along a tree with equal branch lengths (clock-like model) given by the estimate of the chimpanzee branch in brain133 ($t_{\mathrm{H}} = t_{\mathrm{C}} = 0.28$, see Table 3); a fraction *q* of genes evolves along a tree with the same length of the chimpanzee branch as that in the clock-like model ($t_{\mathrm{C}} = 0.28$), while the length of the human branch $t_{\mathrm{Human}} > t_{\mathrm{C}}$ is a free parameter (human-specific model). A justification of this choice for t_{C} is given in the appendix. We then fit the brain133 data in terms of its variance and skewness to this mixture of two models. A general solution to this fitting problem is given by the simple formula $q \cdot (t_{\mathrm{Human}} - t_{\mathrm{C}}) = \text{constant}$ (see appendix). Figure 6A shows the relation of the fraction *q* of genes evolving according to the human-specific model and the length of the human branch t_{Human}. The smaller this fraction *q* is, the longer the human branch of the human-specific evolution tree becomes. With $q = 100\%$ the mixture model is reduced to the single model we estimated in Table 3 ($t_{\mathrm{Human}} = 0.82$). Figure 6B shows, for the case $q = 10\%$, the contribution of 90% of clock-like genes (dashed line) and 10% of genes evolving according to a human-specific model (solid line) to the human-chimpanzee difference distribution. Figure 6C shows how a mixture of the two distributions in Figure 6B can yield a distribution that is indistinguishable from the observed brain133 data. Thus, it is possible that a relatively small number of genes have changed their mode of evolution in the human brain.

Unfortunately, from these data it is not possible to determine the precise number of human-specific gene expression changes or to identify the corresponding genes. Obviously, it is also not clear whether the excess of gene expression changes in the human lineage is due to positive selection or a relaxation of selective constraints. In fact, the determination of the evolutionary mode for the expression of individual genes remains a great challenge. However, we are hopeful that the model presented here can serve as a starting point for such an endeavor by providing a testable null hypothesis that when falsified may indicate the effects of different forms of selection.

We thank Arndt von Haeseler, Ines Hellmann, Michael Lachmann, and Wolfgang Enard for fruitful discussion, technical support, and improving the manuscript. Financial support by the Max Planck Society, the Heinrich-Heine-University, and the Bundesministerium für Bildung und Forschung is gratefully acknowledged.

LITERATURE CITED

Balding, D. J., M. Bishop and C. Cannings, 2003 *Handbook of Statistical Genetics*. John Wiley & Sons, Chichester, UK.

Bolstad, B. M., R. A. Irizarry, M. Astrand and T. P. Speed, 2003 A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics **19:** 185–193.

Caceres, M., J. Lachuer, M. A. Zapala, J. C. Redmond, L. Kudo *et al.*, 2003 Elevated gene expression levels distinguish human from non-human primate brains. Proc. Natl. Acad. Sci. USA **100:** 13030–13035.

Cheung, V. G., L. K. Conlin, T. M. Weber, M. Arcaro, K. Y. Jen *et al.*, 2003 Natural variation in human gene expression assessed in lymphoblastoid cells. Nat. Genet. **33:** 422–425.

Durrett, R., 2002 *Probability Models for DNA Sequence Evolution*. Springer, New York.

Edwards, A. F. W., and L. L. Cavalli-Sforza, 1964 Reconstruction of evolutionary trees, pp. 67–76 in *Phenetic and Phylogenetic Classification*, edited by W. H. Heywood and J. McNeill. Syst. Assoc., London.

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

Feller, W., 1957 *An Introduction to Probability Theory and Its Applications*. John Wiley & Sons/Chapman & Hall, New York/London.

Felsenstein, J., 1973 Maximum-likelihood estimation of evolutionary trees from continuous characters. Am. J. Hum. Genet. **25:** 471–492.

Glazko, G. V., and M. Nei, 2003 Estimation of divergence times for major lineages of primate species. Mol. Biol. Evol. **20:** 424–434.

Gu, J., and X. Gu, 2003 Induced gene expression in human brain after the split from chimpanzee. Trends Genet. **19:** 63–65.

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

Hsieh, W. P., T. M. Chu, R. D. Wolfinger and G. Gibson, 2003 Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles. Genetics **165:** 747–757.

Ihaka, R., and R. Gentleman, 1996 R: a language for data analysis and graphics. J. Comput. Graph. Stat. **5:** 299–314.

Jin, W., R. M. Riley, R. D. Wolfinger, K. P. White, G. Passador-Gurgel *et al.*, 2001 The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. Nat. Genet. **29:** 389–395.

Johnson Norman, L., S. Kotz and N. Balakrishnan, 1995 *Continuous Univariate Distributions*. Wiley, New York/Chichester, UK.

Khaitovich, P., G. Weiss, M. Lachmann, I. Hellmann, W. Enard *et al.*, 2004 A neutral model of transcriptome evolution. PLoS. Biol. **2:** 682–689.

Khaitovich, P., I. Hellmann, W. Enard, K. Nowick, M. Leinweber *et al.*, 2005 Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. Science (in press).

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK.

Lande, R., 1976 Natural-selection and random genetic drift in phenotypic evolution. Evolution **30:** 314–334.

Lynch, M., and W. G. Hill, 1986 Phenotypic evolution by neutral mutation. Evolution **40:** 915–935.

Morley, M., C. M. Molony, T. M. Weber, J. L. Devlin, K. G. Ewens *et al.*, 2004 Genetic analysis of genome-wide variation in human gene expression. Nature **430:** 743–747.

Oleksiak, M. F., G. A. Churchill and D. L. Crawford, 2002 Variation in gene expression within and among natural populations. Nat. Genet. **32:** 261–266.

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

Rise, M. L., K. R. von Schalburg, G. D. Brown, M. A. Mawer, R. H. Devlin *et al.*, 2004 Development and application of a salmonid EST database and cDNA microarray: data mining and interspecific hybridization characteristics. Genome Res. **14:** 478–490.

Schadt, E. E., S. A. Monks, T. A. Drake, A. J. Lusis, N. Che *et al.*, 2003 Genetics of gene expression surveyed in maize, mouse and man. Nature **422:** 297–302.

Su, A. I., M. P. Cooke, K. A. Ching, Y. Hakak, J. R. Walker *et al.*, 2002 Large-scale analysis of the human and mouse transcriptomes. Proc. Natl. Acad. Sci. USA **99:** 4465–4470.

Swofford, D. L., G. J. Olsen, P. J. Waddell and D. M. Hillis, 1996 Phylogenetic inference, pp. 407–514 in *Molecular Systematics*, edited by D. M. Hillis, C. Moritz and B. K. Mable. Sinauer Associates, Sunderland, MA.

Wittkopp, P. J., B. K. Haerum and A. G. Clark, 2004 Evolutionary changes in cis and trans gene regulation. Nature **430:** 85–88.

Communicating editor: L. Excoffier

APPENDIX

Let models A and B describe the evolution of expression differences in terms of our proposed difference of two compound Poisson processes with an asymmetric distribution of mutational effects. The two models differ only in the underlying evolutionary tree: In model A this tree consists of two branches of different lengths t_{Human} and t_{C}; model B defines a tree with branches of equal length t_{C}. Let M describe the mixture of the two models A and B, where the proportions of models A and B are given by q and $1 - q$, respectively. Let Z be the random variable describing expression differences under the respective models. If $f(z|A)$ and $f(z|B)$ define the probability densities of Z under models A and B, the density of Z given M can be computed as

$$f(z|M) = q \cdot f(z|A) + (1 - q) \cdot f(z|B).$$

Using conditional moments the kth moment of Z given M is easily derived:

$$E(Z^k|M) = q \cdot E(Z^k|A) + (1 - q) \cdot E(Z^k|B).$$

Under both models A and B the expectation of Z equals zero. Thus, the expectation of $Z|M$ equals zero and the kth moment of $Z|M$ coincides with its kth central moment.

The variance of $Z|M$ can be computed as

$$\begin{aligned}\mu_2(Z|M) &= q \cdot \mu_2(X) \cdot (t_{\mathrm{Human}} + t_{\mathrm{C}}) + (1 - q) \cdot \mu_2(X) \cdot 2t_{\mathrm{C}} \\ &= \mu_2(X) \cdot (q \cdot (t_{\mathrm{Human}} - t_{\mathrm{C}}) + 2t_{\mathrm{C}}).\end{aligned}$$

Since the skewness of $Z|B$ is equal to zero, the coefficient of skewness of $Z|M$ has the following form:

$$\gamma_1(Z|M) = \frac{q \cdot E(Z^3|A)}{\mu_2(Z|M)^{3/2}} = \frac{q \cdot \gamma_1(X) \cdot (t_{\mathrm{Human}} - t_{\mathrm{C}})}{(q \cdot (t_{\mathrm{Human}} - t_{\mathrm{C}}) + 2t_{\mathrm{C}})^{3/2}}.$$

This set of two equations can be solved explicitly for t_{C}:

$$t_{\mathrm{C}} = \frac{1}{2}\frac{\mu_2(Z|M)}{\mu_2(X)}\left(1 - \frac{\gamma_1(Z|M)}{\gamma_1(X)}\left(\frac{\mu_2(Z|M)}{\mu_2(X)}\right)^{1/2}\right).$$

This solution does not dependent on the parameter q. It also coincides with the smaller solution t_2 in Equation 4, if $\mu_2(X)$ is expressed in terms of fourth moments. Thus, our estimate for t_{C} is the same for all mixtures of this kind and equals the estimate under the simple model. With t_{C} being a constant with respect to the mixture of models A and B, solutions for the remaining parameters q and t_{Human} are easily found as being of the form

$$q \cdot (t_{\mathrm{Human}} - t_{\mathrm{C}}) = \frac{\gamma_1(Z|M)}{\gamma_1(X)}(\mu_2(Z|M))^{3/2} = \text{constant.} \qquad \text{(A1)}$$