Modeling Gene Expression Evolution with an Extended Ornstein–Uhlenbeck Process Accounting for Within-Species Variation

Rori V. Rohlfs,*[1] Patrick Harrigan,[2] and Rasmus Nielsen[1]

[1]Department of Integrative Biology, University of California, Berkeley

[2]Division of Bioinformatics, University of California, San Francisco

***Corresponding author:** E-mail: rrohlfs@berkeley.edu.

Associate editor: Katja Nowick

Abstract

Much of the phenotypic variation observed between even closely related species may be driven by differences in gene expression levels. The current availability of reliable techniques like RNA-Seq, which can quantify expression levels across species, has enabled comparative studies. Ornstein–Uhlenbeck (OU) processes have been proposed to model gene expression evolution as they model both random drift and stabilizing selection and can be extended to model changes in selection regimes. The OU models provide a statistical framework that allows comparisons of specific hypotheses of selective regimes, including random drift, constrained drift, and expression level shifts. In this way, inferences may be made about the mode of selection acting on the expression level of a gene. We augment this model to include within-species expression variance, allowing for modeling of nonevolutionary expression variance that could be caused by individual genetic, environmental, or technical variation. Through simulations, we explore the reliability of parameter estimates and the extent to which different selective regimes can be distinguished using phylogenies of varying size using both the typical OU model and our extended model. We find that if individual variation is not accounted for, nonevolutionary expression variation is often mistaken for strong stabilizing selection. The methods presented in this article are increasingly relevant as comparative expression data becomes more available and researchers turn to expression as a primary evolving phenotype.

Key words: **expression evolution, RNA-Seq, Ornstein–Uhlenbeck, evolutionary models, expression variation.**

Introduction

It has long been posited that gene expression differences explain the bulk of phenotypic diversity across species (King and Wilson 1975). Initial comparative expression analyses based on microarray technology across primates have produced interesting patterns of expression conservation and adaptation and a number of controversial results. For example, studies by Khaitovich et al. (2004a, 2004b) and Gilad et al. (2006) lead to quite different conclusions regarding the importance of natural selection in determining expression level differences and similarities among species. The degree to which selection is acting to modify expression levels is a standing question, especially on the human lineage.

With the advent of reliable technology to quantify gene transcription, the field is now better positioned to explore gene expression evolution and conservation (Gilad, Oshlack, Rifkin 2006; Khaitovich et al. 2006; Whitehead and Crawford 2006; Wang et al. 2009). Specifically, accurate comparative gene expression data is attainable with the developments of both RNA-Seq as a reliable method to quantify expression, and bioinformatical methods to appropriately normalize expression, accounting for species differences (Wang et al. 2009; Trapnell et al. 2010). Several statistical methods have been proposed to investigate the role of natural selection in expression evolution across divergent species by considering expression divergence between species and diversity within species (Hsieh et al. 2003; Rifkin et al. 2003; Nuzhdin et al. 2004; Gilad, Oshlack, Rifkin 2006; Khaitovich et al. 2006; Whitehead and Crawford 2006) or by modeling the expression evolution process (Butler and King 2004; Gu 2004; Oakley et al. 2005; Bedford and Hartl 2008; Blekhman et al. 2008; Chaix et al. 2008; Albert et al. 2012). However, a unified framework has yet to be established, which accounts for the complex variation of gene expression across species, individuals, tissues, environments, and technical replicates. Such sophisticated methods will enable rigorous investigation of the conservation of gene expression, the first step in exploring long-standing hypotheses about the contribution of expression to phenotype (Egger et al. 2004; Kleinjan and van Heyningen 2005; Esteller 2007; Johnstone and Baylin 2010).

A variety of approaches have been used to model gene expression evolution. A number of methods have been implemented considering the ratio of expression divergence to diversity to distinguish expression drift, stabilizing selection, and directional selection (Hsieh et al. 2003; Rifkin et al. 2003; Nuzhdin et al. 2004; Gilad, Oshlack, Rifkin 2006; Khaitovich et al. 2006; Whitehead and Crawford 2006). These nonparametric test statistic approaches are useful to quantify

Article

doi:10.1093/molbev/mst190 Advance Access publication October 10, 2013

divergence and diversity empirically and may provide evidence of different modes of expression evolution but are limited by their inability to formulate complex evolutionary hypotheses or to compare specific models of evolution (Butler and King 2004). A variety of models have been implemented for regression and analysis of variance (ANOVA) analysis, including effect terms for gene, species, individual, microarray probe, interactions between those factors, residual error, and other factors, to explore cases of diverged and conserved expression levels (Rifkin et al. 2003; Gilad, Oshlack, Smyth, et al. 2006; Blekhman et al. 2008; Somel et al. 2009; Blekhman et al. 2010; Warnefors and Eyre-Walker 2012). However, these models typically implicitly assume independence between species and disregard phylogenetic relationships between species, inflating false-positive rates and making them less applicable to complex phylogenies (Felsenstein 1985). In multispecies phylogenies, differing shared evolutionary histories lead to a complex trait covariance structure that can add information and power to evolutionary analyses (Felsenstein 1985).

Phylogenetic structure is taken into account in models of expression evolution based on drift (Brownian motion) processes, allowing for both analysis considering dependencies induced by shared history and more specific formulation and comparison of selective hypotheses (Felsenstein 1985; Butler and King 2004; Gu 2004). Brownian motion processes can effectively model neutral drift (Khaitovich et al. 2005) but are less suitable to model stabilizing selection or conservation, which is expected in the case of gene expression, given simple cellular constraints on expression (Lynch and Hill 1986; Felsenstein 1988).

Ornstein–Uhlenbeck (OU) processes, which model random walks with some pull toward a particular state, have been proposed to model the evolution of quantitative traits subject to both drift and stabilizing selection (Hansen 1997; Butler and King 2004; Bedford and Hartl 2008; Hansen et al. 2008; Kalinka et al. 2010). OU processes include parameters for the degree of drift (σ^2), strength of pull (α), and the particular target value toward which the pull is aimed (θ). In a trait evolution framework, these parameters can be interpreted as phenotype change due to genetic drift, selective force, and optimally fit trait value, respectively, making OU processes a convenient framework in which to investigate selective hypotheses. OU processes have been shown to effectively model gene expression level evolution on divergent phylogenies elucidating the degrees of directional selection at play (Bedford and Hartl 2008; Kalinka et al. 2010; Perry et al. 2012). However, these methods may be limited by their assumption of phylogeny-based variation that does not allow for other sources of variation, for example, environmental, technical, or individual genetic variation (Oakley et al. 2005).

Here, we build upon this work to develop an appropriate statistical model to investigate evolutionary questions using comparative gene expression data with variation across individuals in each species. Gu (2004) alluded to a possible extension of his model that might accomplish this by accounting for "experimental errors" in trait evolution. Ives et al. (2007) proposed a model accounting for "measurement error" across quantitative observations of individuals in a phylogeny. Felsenstein (2008) outlined a model similar to ours based on the work of Lynch (1991), who considered sampling error in trait observations. More recently Hansen and Bartoszek (2012) have formulated an alternate model to account for trait observational and biological variation in evolutionary models. We present an OU model likelihood framework and outline specific hypothesis tests while accounting for phylogenetic relationships between species and variation over individuals within species. This model can be used for cases of nonevolutionary variation, neutral expression drift, stabilizing selection on expression, and lineage-specific shifts in expression level. Our framework builds directly upon that proposed by Bedford and Hartl (2008), which has been used effectively in the literature (Kalinka et al. 2010; Perry et al. 2012), but our model includes a new parameter for within-species variation (τ^2). By modeling within-species variation, we allow the possibility of nonevolutionary expression variation, which is thought to be important in expression (Idaghdour et al. 2010; Pickrell et al. 2010; Price et al. 2011), and improve rigor of distinguishing different regimes of gene expression evolution.

Using the likelihood ratio test presented in this article, Brawand et al. (2011) considered selection on expression in a RNA-Seq data set of ten species, with two to six individuals per species, across six tissues. The analysis accounted for expression variation within species, and tests for shifts in expression levels in each species and branch were performed across the six tissues. This analysis showed that the testis had the largest number of expression shifts, while the brain showed few expression shifts. These results closely mimic those previously found at the DNA level, which suggest that testis-specific genes often are targeted by positive selection, while genes with primary expression in the brain tend to be highly conserved (Nielsen et al. 2005). An exception was the primate lineage, in which the largest number of expression optimum shifts was found in the brain. This could be caused by biological factors such as the evolution of more complex function, or perhaps, alternatively, reflect differences in sampling and treatment of tissues between primates and nonprimates.

Here, we expand upon the applied analysis, describing the method in detail and estimate power and false-positive rates of the tests performed under a variety of circumstances. Specifically, we compare the performance of the model proposed by Bedford and Hartl (2008) (species mean method) to our extended model accounting for variance within species (species variance method). Using simulations, we compare parameter estimation accuracy and ability to distinguish between various selective hypotheses between these two methods. Our results show that a nonevolutionary expression variance model may not be distinguishable from a model of severe stabilizing selection. When using the species mean method that does not allow for a nonevolutionary model, genes that are subject to nonevolutionary environmental variation will often be mistaken as being under intense stabilizing selection. However, we show that the addition of parameter describing within-species variation facilitates statistically valid investigation of nonevolutionary expression variance hypotheses and circumvents the problem of false inferences of strong stabilizing selection when within-species expression variation is large compared with between-species variation. We explore the power of these methods to detect expression shifts in phylogenies, finding the methods to have similar power. In addition to describing the extended species variance model and its power, our results further describe the behavior of the previously published and applied species mean model, which is necessary for rigorous interpretation of results from previous and current studies.

New Approaches

Expression Levels over Individuals within a Species

We implement two methods for modeling expression evolution: the species mean method (as described by Bedford and Hartl [2008]) and the species variance method. In the species mean method, the mean expression level is taken for each species and used as the value of the OU process at that node. In the species variance method, the gene expression levels are assumed to be normally distributed across individuals within a species, with mean given by the underlying OU process value for that species node and variance parameterized by τ^2. This additional parameter models biological and technical variance within species and allows formulation of an additional variance model, the nonevolutionary environmental variance model, as discussed further later.

Investigating Evolutionary Questions

This OU framework can model various specific hypotheses about the nature of gene expression evolution by placing constraint on what values parameters may take at different branches in the phylogeny. The likelihood function for the parameters of the process can be calculated using observed gene expression data for different models so that likelihood ratios can be used to test these models. We propose a series of models and likelihood ratios as a natural starting point to investigate basic questions in comparative analyses of gene expression evolution. We consider models for expression nonevolution, drift, stabilization, and lineage-specific shift. An overview of the nested hypothesis tests based on these models can be seen in table 1.

Table 1. Nested Hypothesis Tests for Evolutionary Models.

	Null Hypothesis	Alternative Hypothesis
Nonevolutionary versus drift	$\sigma^2 = 0, \alpha = 0$	$\sigma^2 > 0, \alpha = 0$
Drift versus stabilization	$\sigma^2 > 0, \alpha = 0$	$\sigma^2 > 0, \alpha \geq 0$
Stabilization versus shift	$\sigma^2 > 0, \alpha \geq 0, \theta_0 = \theta_1$	$\sigma^2 > 0, \alpha \geq 0, \theta_0 \neq \theta_1$

Models

Four models were used to simulate data: 1) the nonevolutionary model where expression does not evolve over time but variation is due to technical, environmental, and individual genetic variation, 2) the neutral drift model where gene expression levels are subject to unconstrained neutral drift over the phylogeny, 3) the stabilizing selection model where gene expression levels drift randomly but are constrained by stabilizing selection, and 4) the selective shift model where expression level experiences stabilizing selection toward different optimally fit expression levels on different branches in the phylogeny, approximating directional (positive) selection in favor of change in expression levels on some lineages of a phylogeny.

If expression levels are not evolving, the observed variance in the data is explained by within-species variation (τ^2) alone. That is, the comparative expression levels can be described by a star phylogeny with zero branch lengths. To construct this model, we eliminate evolutionary drift (set $\sigma^2 = 0$) and stabilizing selection (set $\alpha = 0$) on every branch of the phylogeny. Under this model without phylogenetic signal, gene expression levels are normally distributed across species and individuals as $X_{ik} \sim N(\theta_{\text{root}}, \tau^2)$. Without stabilizing selection, the optimal gene expression value θ is undefined, and we instead estimate the ancestral gene expression value at the root, θ_{root}.

We modeled the case of neutral drift of expression levels by allowing evolutionary drift ($\sigma^2 > 0$) and disallowing stabilizing selection ($\alpha = 0$), which enables covariance of gene expression values between species due to shared evolutionary history. Note that this OU process with $\alpha = 0$ is equivalent to a Brownian motion process of random unconstrained drift.

In the stabilizing selection model, expression levels are subject to drift with a pull toward an optimal (most fit) value. This is modeled by an OU process with $\alpha \geq 0$, where expression evolution is driven toward an optimum expression level θ. Because under the stabilization model, θ is defined, we have some choice in how to model the expression value at the root. As an OU process has the stationary distribution $N(\theta, \frac{\sigma^2}{2\alpha})$, we can chose whether to proceed as in the nonevolving and drift models, including an additional parameter θ_{root}, or to instead use the stationary distribution on this branch. Estimating θ_{root} is equivalent to requiring $\sigma^2 = 0$ on this branch. In this analysis, we use the stationary distribution to describe expression at the root node.

A shift in expression levels can be modeled by allowing θ to vary across the phylogeny. In the shifted expression model, each node i, the expression optimum $\theta_i \in \{\theta_1, \theta_2, \ldots, \theta_n\}$ where n specifies the number of optima hypothesized to act on the phylogeny. Expression at the root node is assumed to follow the OU process stationary distribution $[N(\theta, \frac{\sigma^2}{2\alpha})]$ as in the stabilization model.

Simulating Expression Data

Using the models listed earlier, for a variety of parameter values, we determined the expression level distributions and simulated comparative expression data sets. For the phylogenetic structure, we used a ten-leaf phylogeny with two to six individuals per species, equivalent to that considered by Brawand et al. (2011), which we refer to here as "small tree." To explore the effect of phylogeny size and sample size per species on power, we consider two additional phylogenetic structures: a "deep tree," which is constructed with four copies of the small tree st as [(st,st),(st,st)] with all connecting branches the length of st itself, and a "wide tree," which is the same phylogenetic structure as the small tree, but with four times the individuals in each species (supplementary fig. S1, Supplementary Material online). For each method, power and false positive rate data points are shown based on 1,000 simulated replicates. The specific parameter values used for each simulation vary across models and are indicated on all figures.

Results

The Effect of Ignoring Individual Variation

The commonly used OU-based model for expression evolution does not directly account for measurement of expression levels in multiple individuals of the same species. Typically, when these models are used to calculate the probability of observed data, the sample means are used as species expression levels, a technique we refer to as the species mean method. As an alternative, we propose the species variance method, where within-species variation is taken into account with an additional parameter, τ^2. Note that the species mean method approximates the species variance method as the number of individuals per species increases. Using both the species mean and variance methods, where possible, we compute likelihood ratios to distinguish expression subject to nonevolutionary variance, drift, stabilization, and lineage-specific shifts (table 1).

Test 1: Testing for Phylogenetic Signal

Because the mean species method does not allow for within-species variation, in the absence of selection, any variation in mean expression levels between species must be explained by drift. In this way, under the species mean model, data simulated with any kind of variation between species is always more likely under the drift model ($\sigma^2 > 0$) than the nonevolution model ($\sigma^2 = 0$). As an illustration, data were simulated with the species variance method under the nonevolutionary model and its likelihood computed for various values of σ^2. Figure 1 shows this likelihood surface with the species mean method to have a peak at $\sigma^2 > 0$, indicating that, even for data simulated without evolutionary information, the species mean method will assign some phylogenetic signal. Any attempt to perform the test for phylogenetic signal under the species mean model will result in rejection of the null hypothesis, meaning both power and false-positive rate are 1.0.

Because the species variance method considers individual variation, it allows for species mean expression levels to vary even in the absence of evolutionary drift. In other words, the species variance method enables a nonevolutionary model for gene expression variance. So, with the species variance method, the probabilities of the observed individual expression levels can be compared under a model of nonevolutionary variation and a model of evolutionary drift (table 1). Distinguishing evolutionary and nonevolutionary variation is particularly relevant because the expression of many genes is thought to be subject to intense environmental variation (Idaghdour et al. 2010).

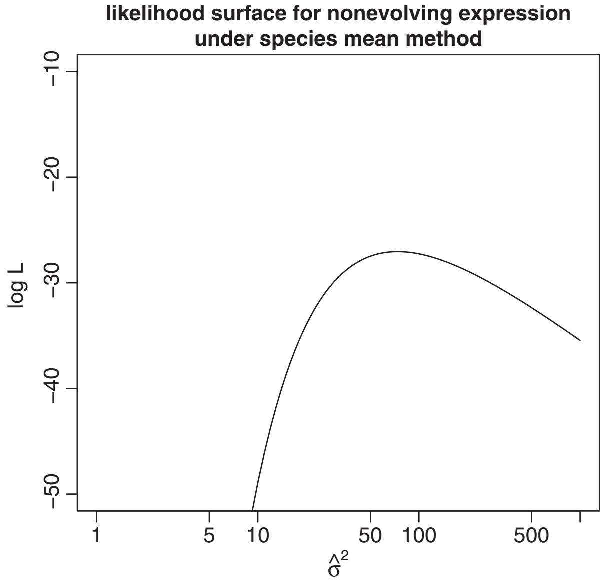

Fig. 1. The species mean model log likelihood function for data simulated under the nonevolutionary species variance model with $\tau^2 = 5$ (within-species variation) is computed with $\hat{\sigma}^2$ (estimated drift) fixed and other parameters optimized (θ_{root} [estimated expression at root] and $\hat{\tau}^2$ [estimated within-species variation]).

Gene expression levels were simulated over the three phylogenetic structures under the nonevolutionary model with varying values of τ^2 and under the drift model with varying values of both τ^2 and σ^2. Using the species variance method, the likelihood ratio for nonevolution versus drift was computed for all these data, enabling simulation-based estimates of power and false-positive rates, as shown in figure 2. The critical value for hypothesis rejection was determined using the simulations under the null hypothesis to attain a nominal false-positive rate of 0.05. Because the nonevolution and drift models are simply nested and differ by one parameter, bounded at zero, the expected asymptotic distribution of the likelihood ratio test statistic is a 50:50 mixture of chi-square with one degree of freedom and a point mass at zero. As expected, power increases with drift (parameterized by σ^2), decreases with within-species variation (parameterized by τ^2), and is greater for larger phylogenies.

Test 2: Testing for Stabilizing Selection

The likelihood function under both the drift and stabilizing selection models is fully computable using either the species mean or variance method. The resulting likelihood ratio can be used to distinguish data simulated under each model. This likelihood ratio was computed for data simulated under the nonevolutionary, drift, and stabilization models using a variety of parameter values. Figure 3 shows the positive identification

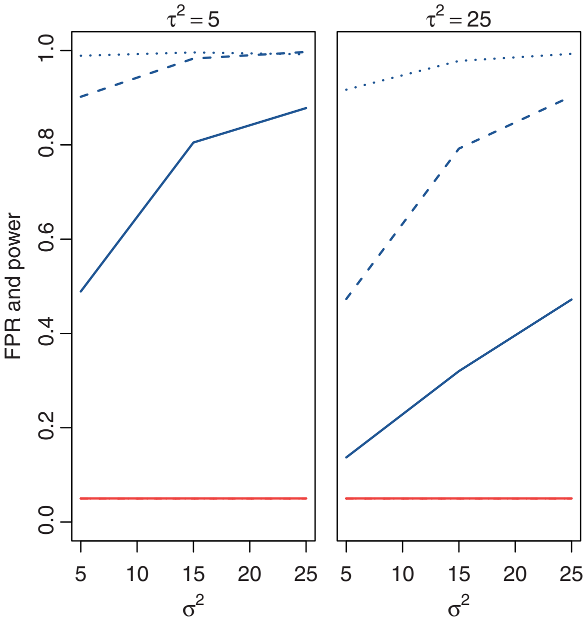

Fig. 2. No evolution versus drift test (test 1) false-positive rate (lower curves) and power (upper curves) using the species variance model for various simulated values of τ^2 (within-species variation) and σ^2 (drift) using critical values for a nominal false-positive rate of 0.05, run with different phylogenetic structures (solid: small tree, dashed: wide tree, dotted: deep tree).

rates of data simulated under drift (false-positive rate) and stabilization (power) using the species mean, species variance, and conditioned species variance methods. In the conditioned species variance method, rejection of the null hypothesis is conditioned on rejection of the nonevolutionary model as well. That is, a gene must be shown to be undergoing expression evolution according to the known phylogeny before further tests about the mode of expression evolution are performed. In this case, the nonevolutionary model must be rejected in favor of the drift model before the test for stabilizing selection is performed. So stabilizing selection is identified when both the nonevolutionary versus drift and the drift versus stabilization tests reject the null. Power is higher for larger phylogenies and using the species mean method, as opposed to the species variation models.

However, data simulated under the nonevolutionary model is often misidentified as under stabilizing selection using the species mean method, as shown in table 2 and figure 3. This can be explained by a lack of identifiability when distinguishing between the nonevolutionary and stabilization models. In the limit of strong selection, all species expression levels will take the same (optimal) value. This is equivalent to the nonevolutionary model used here, where variation is not phylogenetic but individual. In the species mean method, within-species or sampling variation is not modeled, and similar expression mean values across species are better explained by intense stabilizing selection than by drift or nonevolution. As a result of using the species mean method, genes with no phylogenetic signal often appear to be under stabilizing selection, yielding high false-positive rates (table 2).

Because the critical values for these tests are chosen to attain a null hypothesis false-positive rate of 0.05, critical values vary over phylogenetic structures and methods (supplementary table S1, Supplementary Material online). Specifically, because the deep tree contains more information that can be exploited with the species variance method, this configuration has more power to distinguish expression under drift and stabilization. In this case, data simulated under drift have very low likelihood ratio values, so the critical threshold is remarkably low to attain a false-positive rate as high as 0.05. When considering truly nonevolutionary expression in the drift versus stabilization test, this results in an elevated false-positive rate.

Using the species mean method, power increases with strength of stabilizing selection (α), while using the species variation method, power decreases with α. Again, consider the limit of expression under the intense stabilizing selection that erases information about ancestral expression levels in observations of extant species so that there is little to no variation in expression level between species. When using the species variation method, as α increases, the data may be better explained by parameter values resembling the nonevolutionary model than the stabilization model. So the test for stabilizing selection may actually lose power under intense stabilizing selection, because in our construction, it entertains a lack of phylogenetic signal as one of the possible alternatives.

Test 3: Testing for Expression Level Shifts

Likelihood ratios comparing the stabilization and shift models were computed for data simulated under the stabilization and shift model with varying distances between the two optima ($\Delta\theta$). Again, critical values were chosen to achieve a nominal false-positive rate of 0.05. Figure 4 shows the power of this test for different phylogenies, values of $\Delta\theta$, and methods. Power increases with phylogeny size and $\Delta\theta$. In these simulations, the species variance method has higher power than the species mean method, but the power is reduced with the conditioned species variance method. Because data simulated under selective shift depart from the phylogenetic structure, according to the magnitude of shift, the conditioned species variance method loses some power (supplementary table S2, Supplementary Material online).

The expression shift model produces patterns of expression levels that depart from those under the nonevolving, drift, and stabilizing models. Data simulated under both the nonevolving and drift models are rarely misidentified as a product of the expression shift model, as seen in tables 3 and 4.

Estimating Model Parameters

For each of the four expression evolution models and the two individual variation methods the likelihood function is optimized to provide joint maximum likelihood estimates of all the parameters. The nonevolutionary model is parameterized in terms of the ancestral expression level (θ_{root}) and the within-species variance (τ^2); the drift model by θ_{root}, τ^2, and strength of drift (σ^2); the stabilizing selection model by the

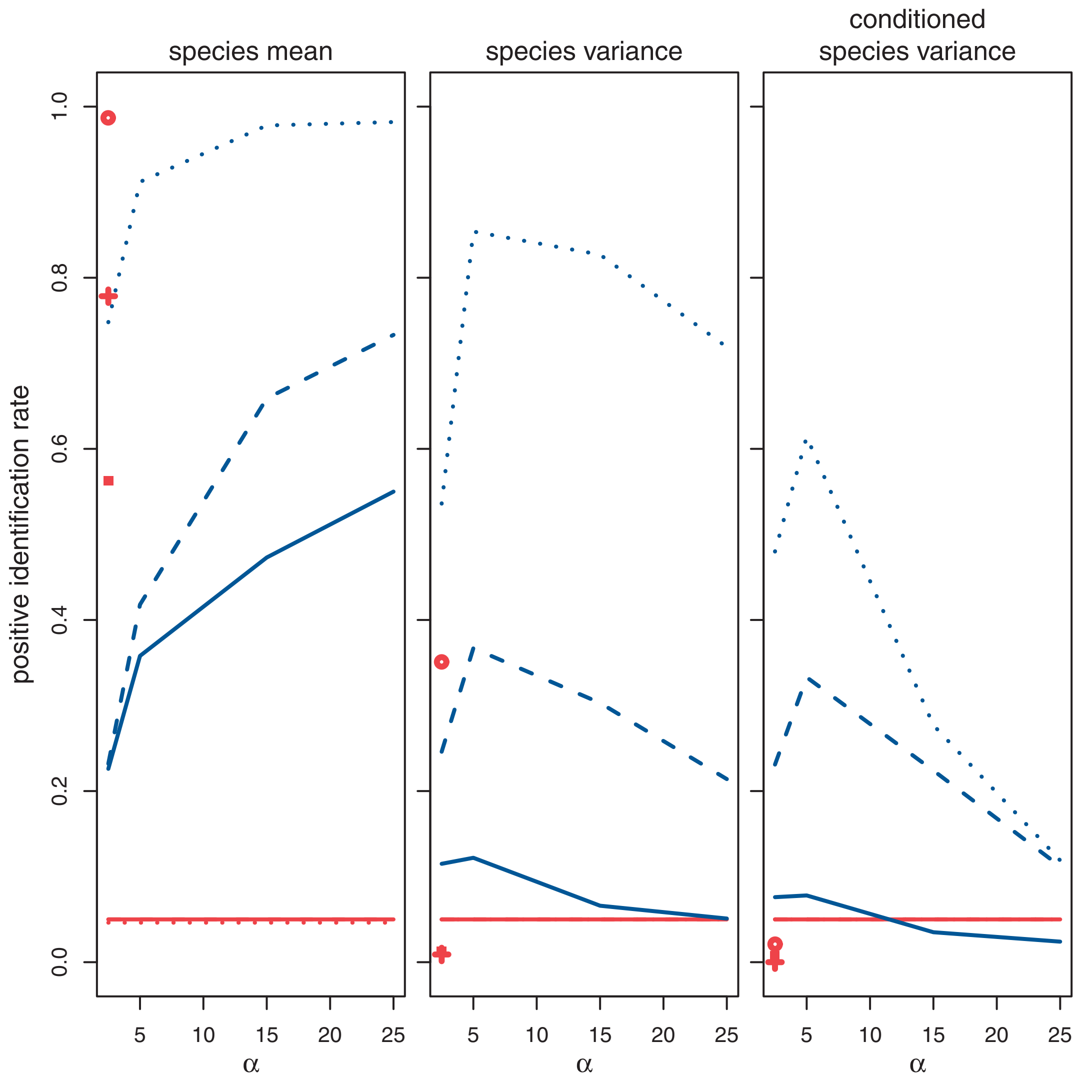

FIG. 3. Drift versus stabilization test (test 2) false-positive rate (lower curves) and power (upper curves) using the species mean, species variance, and conditioned species variance models for various simulated values of α (strength of pull) using critical values for a nominal false-positive rate of 0.05, run with different phylogenetic structures (solid: small tree, dashed: wide tree, dotted: deep tree). Rates for data simulated under the nonevolutionary model and misidentified as under stabilizing selection are shown as points on the left of each plot (square: small tree, cross: wide tree, circle: deep tree).

Table 2. False-Positive Rate of Truly Nonevolutionary Data in Drift versus Stabilization Test.

	Typical Tree	Deep Tree	Wide Tree
Species mean	0.56	0.99	0.78
Species variance	0.01	0.35	0.01
Conditioned species variance	0.01	0.02	0.00

optimal expression level (θ), τ^2, σ^2, and the strength of stabilizing selection (α); and the selective shift model by the different optimal expression levels defined on specific branches (θ_1, θ_2 for two optima) τ^2, σ^2, and α. Note that the within-species variation parameter τ^2 is undefined in the species mean method.

In each case, we simulate data under the true model to assess best-case parameter estimation accuracy. The parameters for ancestral expression (θ_{root}) and within-species variance (τ^2) enter into the likelihood linearly as the mean and variance of expression under the nonevolutionary model, enabling accurate estimation with the species variance method (supplementary figs. S2 and S3, Supplementary Material online). Using the species mean model, likelihoods, and therefore parameter estimates, are not computable under a nonevolutionary model. Even in the more complex shift model, the expression level optima parameter estimates ($\hat{\theta}_1$ and $\hat{\theta}_2$) are easily computed as the means of multivariate normal distributions (supplementary fig. S7, Supplementary Material online). In the case of expression shifts on shorter branches, the expression levels may not have had time to reach the optima, leading to reasonable underestimates of these values.

In the models where the parameters for drift (σ^2) and stabilizing selection (α) are defined (drift, stabilization, shift), these parameters enter in a more complex way, making them more difficult to estimate with high accuracy (supplementary figs. S4–S6, Supplementary Material online). Nonetheless, it is

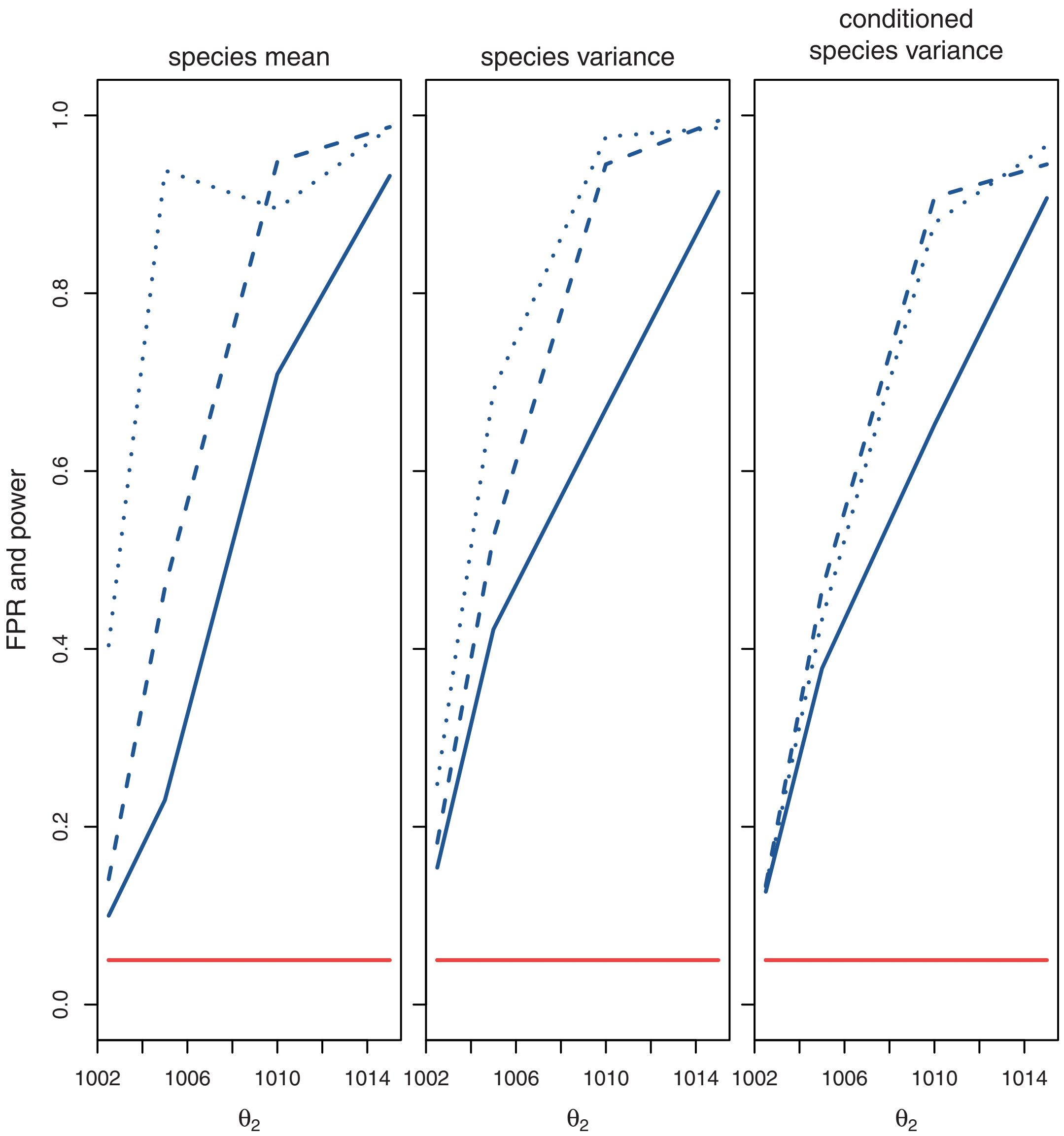

Fig. 4. Stabilization versus expression shift test (test 3) false-positive rate (lower curves) and power (upper curves) using the species mean, species variance, and conditioned species variation model for various simulated values of $\Delta\theta$ (change in expression optimum) using critical values for a nominal false-positive rate of 0.05, run with different phylogenetic structures (solid: small tree, dashed: wide tree, dotted: deep tree).

Table 3. False-Positive Rate of Truly Nonevolutionary Data in Stabilization versus Shift Test.

	Typical Tree	Deep Tree	Wide Tree
Species mean	0.02	0.01	0.02
Species variance	0.01	0.00	0.00
Conditioned species variance	0.01	0.00	0.00

Table 4. False-Positive Rate of Truly Drifting Data in Test of Stabilization versus Expression Shift.

	Typical Tree	Deep Tree	Wide Tree
Species mean	0.07	0.03	0.05
Species variance	0.07	0.02	0.06
Conditioned species variance	0.07	0.02	0.06

interesting to consider how these estimates differ under the species variance and mean methods. Because the species mean method attributes all variation between species expression levels to drift, this method often overestimates $\widehat{\sigma}^2$ when compared with estimates under the species variance method (supplementary figs. S4 and S5, Supplementary Material online). In part to compensate for this overestimation, under the stabilization model, the species mean method has a tendency to overestimate α as well, when compared with the species variation method which is more likely to underestimate α (supplementary fig. S6, Supplementary Material online).

Discussion

We have extended previous methods (Butler and King 2004; Bedford and Hartl 2008) to model gene expression evolution, including a term to account for within-species variation over individuals caused by biological, technical, and environmental inputs. Other studies have shown that RNA-Seq can accurately quantify gene expression levels, but that there can be substantial technical (Marioni et al. 2008; Mortazavi et al. 2008) and biological (Idaghdour et al. 2010; Pickrell et al. 2010; Price et al. 2011) variance. Through simulations, we have shown that this extended model can be used to more accurately infer underlying evolutionary mechanisms.

The parameters of this gene expression evolution model can be estimated using maximum likelihood procedures. In simulations, when considering the correct evolutionary model, some the species expression levels (θ) and within-species variation (τ^2) can be estimated with some accuracy (supplementary figs. S2, S3, and S7, Supplementary Material online). However, the parameter values for the strength of drift (σ^2) and stabilization (α) are dependent on each other, so their individual estimates are less reliable (supplementary figs. S4–S6, Supplementary Material online). Although we do not recommend interpreting the parameter estimates strongly in and of themselves, comparison of likelihood values between models can be used for model choice.

The first test for phylogenetic signal is only possible using the species variance method. In the species mean method, any difference between species expression levels is explained by drift, so the likelihood for realistic (nonpoint mass) data peaks at $\hat{\sigma}^2 > 0$ (fig. 1) and the null hypothesis of nonevolutionary variation is always rejected, except if all individuals have the exact same expression level. The ability to perform this test is important because the expression of many genes appears to be subject to much environmental or individual variation, obscuring phylogenetic signal (Idaghdour et al. 2010; Pickrell et al. 2010; Price et al. 2011). Distinguishing these genes before investigating other hypotheses of selection provides a basic filter for nonevolutionary variation. Using the species variance method, power to detect phylogenetic signal increases with phylogeny size and strength of drift (σ^2) and decreases with within-species variation (τ^2) and can easily be controlled for a desired nominal false-positive rate (fig. 2).

The test for stabilizing selection versus neutral drift is possible using both the species mean and variation methods. The species mean method shows higher power than the species variance method (fig. 3) with critical values chosen to achieve false-positive rates of 0.05 under the null hypothesis of neutral drift. However, the species mean method also suffers from false-positive rates as high as 0.99 for expression levels that are truly nonevolving (table 2). This dramatically elevated false-positive rate renders the species mean method ineffective for distinguishing stabilizing selection. The species variance method has lower false-positive rates for truly nonevolving gene expression levels, though some false-positive rates are still uncontrolled. The false-positive rate can be controlled using the conditioned species variance method, but substantial power is lost as well. This problem of identifiability is explained by the fact that expression levels may not vary much among species under both extreme stabilizing selection and the nonevolving model. This presents an identifiability problem that results in lower power to identify stabilizing selection, even using the species variance method, especially for phylogenies the size of those currently available.

With this reduced power, the experimental results of the species variance method test for stabilizing selection, like those published by Brawand et al. (2011), are not easily interpretable. Robust analysis of stabilizing selection awaits larger data sets. Interestingly, the similarity of expression profiles of genes with nonevolutionary variance and genes under stabilizing selection may partially explain the results of a previous study where a nonevolutionary model was not rejected in favor of a drift model (Oakley et al. 2005).

A number of studies have been published claiming to show widespread conservation and stabilizing selection of gene expression levels. These studies generally either perform an ANOVA-style analysis where expression diversity within species is compared with divergence between species without regard to phylogeny (Lemos et al. 2005; Gilad, Oshlack, Smyth, et al. 2006; Staubach et al. 2010; Schroder et al. 2012; Warnefors and Eyre-Walker 2012) or the studies use an OU model like the species mean method without regard to within-species expression diversity (Bedford and Hartl 2008; Kalinka et al. 2010). Specifically, the species mean method that has been used to support claims of 80% of genes in a set of six *Drosophila* species across developmental time points are under stabilizing selection (Kalinka et al. 2010). For comparison, in our simulations for a ten-species phylogeny using the species mean method, we see 56% of genes with nonevolutionary expression variance are misidentified as being under stabilizing selection. Further, our power to identify stabilizing selection ranges from 22% to 55%. The specific false-positive rates and power depend on parameter values like within-species variance, strength of drift, and the strength of stabilization. We do not contest that stabilizing selection is crucial to functional expression level. However, because nonevolutionary variation is often mistaken as stabilizing selection and the power to identify stabilizing selection may be low, the precise degree to which stabilizing selection versus nonevolutionary variation cannot be distinguished using the species mean method. This results in a potential overstatement of the importance of stabilizing selection in expression evolution. As an increasing number of studies seem to reject neutral expression evolution in favor of stabilizing expression, it will be important to consider the possibility of nonevolutionary inputs to expression level.

The test for expression level shift, which was most prominently featured by Brawand et al. (2011), shows similar high power under both the species mean and species variance models, which is consistent with their strong interpretation of those results.

It is worth noting the interplay between statistical modeling and experimental design. Small organisms are commonly pooled before they are typed for expression level. Typing pooled samples effectively performs an "experimental mean" on expression level, which may reduce individual biological variance when compared with single individual samples. However, pooling samples may not reduce technical variance between typing runs, so modeling this variance is still important. Additionally, the implicit experimental mean may be more robust to biological variance, but it is subject to error and the true species mean is still unknown. As the species variance method allows for error in estimated species mean expression level and therefore allows for nonevolutionary expression differences, the species variance method is still preferable with pooled samples.

Although OU models provide a simple model of gene expression evolution with easily tested hypotheses about selective regimes, much remains to be explored about the mechanism and nature of gene expression evolution. The null hypothesis of nonevolution considered here allows expression levels for all individuals across species to be drawn from the same underlying normal distribution. Other possible models of nonevolution could allow species expression levels to vary in a nonphylogenetic manner, for example, according to different environmental conditions, which may more accurately represent the expression of genes highly influenced by environmental factors. In addition, the OU model implicitly assumes that the effect size of expression mutations follow a normal distribution (parameterized by σ^2). We have much to learn about the mechanisms of expression evolution. Mutation effect sizes may follow a Poisson distribution (Khaitovich et al. 2005) or some mixed model with common small effect sizes and rare large effect mutations (Chaix et al. 2008; Gruber et al. 2012).

As with other models of gene expression evolution, here we have considered a single gene's expression level across species and individuals. The power we estimated is accurate for each marginal single gene's expression levels. Of course, in typical data sets, expression levels are quantified for many genes simultaneously. Because expression levels across genes vary in response to each other over evolutionary time and environmental conditions, biological expression levels are not independent. When accounting for multiple testing across genes, assuming independence may lead to a loss in power. Complex correlations across genes must be considered simultaneously to rigorously understand the biological basis underpinned by full genetic architecture. A rigorous multigene expression evolution analysis awaits development of methods for correlated trait evolution based on previously described models (Lande and Arnold, 1983; Felsenstein, 1985, 1988; Lynch, 1991) that would increase information and power.

The simulations presented here indicate that these methods may be used to distinguish some regimes of gene expression evolution, particularly expression level shifts. However, some expression models, particularly nonevolutionary variance and stabilizing selection, result in similar patterns of expression levels, which are not distinguishable with currently available comparative expression data sets. As more extensive comparative expression data becomes available and the mechanisms of expression variation and evolution are better understood, increasingly appropriate models can be developed to explore hypotheses of gene expression evolution.

Materials and Methods

We model the evolution of a gene's expression level over time as an OU process, which is defined by the stochastic differential equation

$$dX_t = \alpha(\theta - X_t)dt + \sigma dW_t, \quad (1)$$

where X_t is the process value at time t, W_t is a normally distributed random variable with variance dt (W_t $N(0,dt)$), α parameterizes the strength of pull toward the optimal value θ, and σ parameterizes the strength of drift. The change in expression level (dX_t) over interval dt is the sum of stochastic and deterministic components. The stochastic component (σdW_t) is a normally distributed random variate with variance $\sigma^2 dt$, and the deterministic component ($\alpha(\theta - X_t)dt$) describes pull of the process toward θ. In modeling gene expression evolution, the stochastic component of change represents neutral drift in expression level and the deterministic component of change represents stabilizing selection.

The OU Process as a Model Gene Expression Evolution

For comparative analysis, we assume a phylogeny of known topology and branch lengths and assign an OU process to each branch. Formally, for every node i with expression level X_i, we assign the parameters α_i, σ_i^2, and θ_i to the branch leading to that node. Each node expression level X_i is distributed normally with

$$E[X_i] = E[X_p]e^{-\alpha_i t_{ip}} + \theta_i(1 - e^{-\alpha_i t_{ip}}), \quad (2)$$

$$\mathrm{Var}[X_i] = \frac{\sigma_i^2}{2\alpha_i}(1 - e^{-2\alpha_i t_{ip}}) + \mathrm{Var}[X_p]e^{-2\alpha_i t_{ip}}, \quad (3)$$

where X_p is the expression level at the parental node p and t_{ip} is the length of the branch separating i from its parent p. Each of these moments contains a contribution from the ancestral gene expression level that decays at a rate given by the strength of stabilizing selection α. For any two nodes i and j

$$\mathrm{Cov}[X_i, X_j] = \mathrm{Var}[X_a]\exp(-\sum_{k\in l_{ij}} \alpha_k t_k - \sum_{k\in l_{ji}} \alpha_k t_k), \quad (4)$$

where X_a is the expression level at the most recent common ancestor of i and j and l_{ij} denotes the set of all nodes in the lineage of X_i not in the lineage of X_j. Similar to equations (2) and (3), the covariance of any two nodes is determined by the variance at the common ancestor and decays exponentially over the time the nodes have evolved independently since divergence, with a rate given by the strength of stabilizing selection.

The states of the OU processes at the terminal taxa (i.e., species expression levels) are distributed as a multivariate normal as described in equations (2)–(4). The likelihood function under such a specified OU model can be easily computed, enabling the use of maximum likelihood methods to estimate parameter values. Similarly, given a phylogeny and parameter values, the distribution of expression levels obtained according to the multivariate normal and data can be simulated.

Expression Levels in Individuals

In the species variance method, τ^2 confounds all sources of individual variance (e.g., technical, environmental) into one parameter. Formally, at any species node i and individual k, $X_{ik} \sim N(X_i, \tau^2)$, so that $E[X_{ik}] = E[X_i]$, $\mathrm{Var}[X_{ik}] = \mathrm{Var}[X_i] + \tau^2$, and $\mathrm{Cov}[X_{ik}, X_{jl}] = \mathrm{Cov}[X_i, X_j]$ where $i \neq j$.

Supplementary Material

Supplementary figures S1–S7 and tables S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors are immensely grateful to the individuals whose tissue samples were used in this study, without which none of this work would be possible. The authors also thank Mehmet Somel for his thoughtful discussions and anonymous reviewers for their helpful suggestions. This work was supported in part by National Institutes of Health grant 2R14003229-07 and National Science Foundation award 1103767. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Albert F, Somel M, Carneiro M, et al. (14 co-authors). 2012. A comparison of brain gene expression levels in domesticated and wild animals. *PLoS Genet.* 8:e1002962.

Bedford T, Hartl D. 2008. Optimization of gene expression by natural selection. *Proc Natl Acad Sci U S A.* 106:1133–1138.

Blekhman R, Marioni J, Zumbo P, Stephens M, Gilad Y. 2010. Sex-specific and lineage-specific alternative splicing in primates. *Genome Res.* 20: 180–189.

Blekhman R, Oshlack A, Chabot A, Smyth G, Gilad Y. 2008. Gene regulation in primates evolves under tissue-specific selection pressures. *PLoS Genet.* 4:e1000271.

Brawand D, Soumillon M, Necsulea A, et al. (18 co-authors). 2011. The evolution of gene expression levels in mammalian organs. *Nature* 478:343–348.

Butler M, King A. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *Am Nat.* 164:683–695.

Chaix R, Somel M, Kreil D, Khaitovich P, Lunter G. 2008. Evolution of primate gene expression: drift and corrective sweeps? *Genetics* 180: 1379–1389.

Egger G, Liang G, Aparicio A, Jones P. 2004. Epigenetics in human disease and prospects for epigenetic therapy. *Nature* 429: 457–463.

Esteller M. 2007. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat Rev Genet.* 8:286–298.

Felsenstein J. 1985. Phylogenies and the comparative method. *Am Nat.* 125:1–15.

Felsenstein J. 1988. Phylogenies and quantitative characters. *Annu Rev Ecol Syst.* 19:445–471.

Felsenstein J. 2008. Comparative methods with sampling error and within-species variation: contrasts revisited and revised. *Am Nat.* 171:713–725.

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

Gilad Y, Oshlack A, Smyth GK, Speed TP, White KP. 2006. Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature* 440:242–245.

Gruber J, Vogel K, Kalay G, Wittkipp P. 2012. Contrasting properties of gene-specific regulatory, coding, and copy number mutations in *Saccromyces cerevisiae*: frequency, effects and dominance. *PLOS Genet.* 8:e1002497.

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

Hansen T. 1997. Stabilizing selection and the comparative analysis of adaptation. *Evolution* 51:1341–1351.

Hansen T, Bartoszek K. 2012. Interpreting the evolutionary regressions: the interplay between observational and biological errors in phylogenetic comparative studies. *Syst Biol.* 61:413–425.

Hansen T, Pienaar J, Orzack S. 2008. A comparative method for studying adaptation to a randomly evolving environment. *Evolution* 62: 1965–1977.

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

Idaghdour Y, Czika W, Shianna K, et al. (11 co-authors). 2010. Geographical genomics of human leukocyte gene expression variation in southern Morocco. *Nat Genet.* 42:62–67.

Ives A, Midford P, Garland T. 2007. Within-species variation and measurement error in phylogenetic comparative methods. *Syst Biol.* 56: 252–270.

Johnstone S, Baylin S. 2010. Stress and the epigenetic landscape: a link to the pathobiology of human diseases? *Nat Rev Genet.* 11:806–812.

Kalinka A, Varga K, Gerrard D, Preibisch S, Corcoran D, Jarrells J, Ohler U, Bergman C, Tomancak P. 2010. Gene expression divergence recapitulates the developmental hourglass model. *Nature* 468:811–816.

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

Khaitovich P, Pääbo S, Weiss G. 2005. Towards a neutral evolutionary model of gene expression. *Genetics* 170:929–939.

Khaitovich P, Weiss G, Lachmann M, Hellmann I, Enard W, Muetzel B, Wirkner U, Ansorge W, Pääbo S. 2004a. A neutral model of transcriptome evolution. *PLoS Biol.* 2:e132.

Khaitovich P, Muetzel B, She X, et al. (15 co-authors). 2004b. Regional patterns of gene expression in human and chimpanzee brains. *Genome Res.* 14:1462–1473.

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

Kleinjan D, van Heyningen V. 2005. Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am J Hum Genet.* 76:8–32.

Lande R, Arnold SJ. 1983. Measurement of selection on correlated characters. *Evolution* 37:1210–1226.

Lemos B, Meiklejohn CD, Cceres M, Hartl DL. 2005. Rates of divergence in gene expression profiles of primates, mice, and flies: stabilizing selection and variability among functional categories. *Evolution* 59: 126–137.

Lynch M. 1991. Methods for the analysis of comparative data in evolutionary biology. *Evolution* 45:1065–1080.

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

Marioni J, Mason C, Mane S, Stephens M, Gilad Y. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* 18:1509–1517.

Mortazavi A, Williams B, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Methods* 5:621–628.

Nielsen R, Bustamante C, Clark A, et al. (13 co-authors). 2005. Molecular signatures of natural selection. *PLoS Biol.* 3:e170.

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

Oakley T, Gu Z, Abouheif E, Patel N, Li W. 2005. Comparative methods for the analysis of gene-expression evolution: an example of using yeast functional genomic data. *Mol Biol Evol.* 22:40–50.

Perry G, Melsted P, Marioni J, et al. (12 co-authors). 2012. Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome Res.* 22:602–610.

Pickrell J, Marioni J, Pai A, Degner J, Engelhardt B, Nkadori E, Veyrieras J-B, Stephens M, Gilad Y, Pritchard J. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464:768–772.

Price A, Helgason A, Thorleifsson G, McCarroll S, Kong A, Stefansson K. 2011. Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genet.* 7: e1001317.

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

Schroder K, Irvine KM, Taylor MS, et al. (25 co-authors). 2012. Conservation and divergence in toll-like receptor 4-regulated gene expression in primary human versus mouse macrophages. *Proc Natl Acad Sci U S A.* 109:E944–E953.

Somel M, Franz H, Yan Z, et al. (15 co-authors). 2009. Transcriptional neoteny in the human brain. *Proc Natl Acad Sci U S A.* 106: 5743–5748.

Staubach F, Teschke M, Voolstra CR, Wolf JB, Tautz D. 2010. A test of the neutral model of expression change in natural populations of house mouse subspecies. *Evolution* 64:549–560.

Trapnell C, Williams B, Pertea G, Mortazavi A, Kwan G, van Baren M, Salzberg S, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol.* 28: 511–515.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet.* 10:57–63.

Warnefors M, Eyre-Walker A. 2012. A selection index for gene expression evolution and its application to the divergence between humans and chimpanzees. *PLoS One* 7:e34935.

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