Evaluating the Performance of Widely Used Phylogenetic Models for Gene Expression Evolution

Jose Rafael Dimayacyac[1], Shanyun Wu[2], and Matt Pennell[3,4,5,*]

[1]Michael Smith Laboratories, University of British Columbia, Vancouver, Canada

[2]Division of Biology and Biomeidcal Sciences, Washington University in St. Louis, St. Louis, USA

[3]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, USA

[4]Department of Biological Sciences, University of Southern California, Los Angeles, USA

[5]Department of Zoology, University of British Columbia, Vancouver, Canada

*Corresponding Author: mpennell@usc.edu

Abstract

Phylogenetic comparative methods allow biologists to make inferences about the evolutionary history of phenotypes. These methods are increasingly used to study the evolution of gene expression. However, it is unknown whether the distributional assumptions of phylogenetic models designed for quantitative phenotypic traits are realistic for expression data (i.e., how well do the models actually perform?); and the reliability of conclusions of phylogenetic comparative studies of gene expression may depend on whether the data is well-described by the chosen model. To evaluate this, we first fit several phylogenetic models of trait evolution to 9 previously published comparative expression datasets, comprising a total of 54,774 genes with 155,679 unique gene-tissue combinations. Using a previously developed approach, we then assessed how well the best model of the set described the data in an absolute (not just relative) sense. First, we find that Ornstein-Uhlenbeck models were the preferred model for 59.8% of gene-tissue combinations. Second, we find that for 39% of gene-tissue combinations, the best fit model was found to perform poorly by at least one of the test statistics we examined. Third, we find that when simple models do not perform well, this appears to be typically a consequence of failing to fully account for heterogeneity in the rate of the evolution of gene expression across lineages. We advocate that assessment of model performance should become a routine component of phylogenetic comparative expression studies; doing so can improve the reliability of inferences and inspire the development of novel models.

Introduction

While DNA holds the genetic information required for life to work, other elements are largely required for cells to function. These functional elements are responsible for the molecular processes that eventually lead to phenotypes (Kellis et al. 2014). The most prominently studied of these elements is gene expression. There is a long tradition of thinking about gene expression evolution in a comparative context (King & Wilson 1975), yet it is only recently that it has been feasible to gather transcriptomic data for multiple species in a standardized way – this has opened new avenues for investigating the evolutionary processes responsible for generating diversity (Hill et al. 2021; Price et al. 2022) of changes in gene expression.

Identifying interspecies differences in gene expression can pinpoint which sets of genes are responsible for differences between organisms. For example, Chen et al. 2021 recently investigated the expression of the ACE2 receptor across species and cell types to identify susceptibility of different mammals to SARS-CoV-2, where species with higher expression of this receptor in respiratory cells were deemed to be at a higher risk (Chen et al. 2021). Many such studies have used the approach of directly comparing gene expression levels between orthologs to understand an array of topics, such as the function of epigenetic modifications (Cain et al. 2011), the connection between DNA and methylation (Hernando-Herraez et al. 2015), and the evolution of enhancer regions (Villar et al. 2015). The studies mentioned above (in addition to many others in the field) use pairwise comparisons in which all gene expression values from all species are compared to one another. Essentially, this assumes that gene expression values from different species all represent independent measurements (Dunn et al. 2018). However, due to their shared evolutionary history, more closely related species will resemble each other in many ways and some of these shared (and, in many cases, unmeasured) attributes will influence how focal variables (here, gene expression and some attribute of interest) are associated with one another (Felsenstein 1985; Uyeda et al. 2018). While this challenge has been widely recognized across the biological sciences, many comparative gene expression studies still do multi-species comparisons with sequential pairwise comparisons, which a recent study demonstrated could be highly misleading (Dunn et al. 2018).

In addition to controlling for unobserved (and phylogenetically structured) confounding variables, phylogenetic comparative methods (**PCMs**; for recent reviews of these methods see Pennell & Harmon 2013; Garamszegi 2014; and Harmon 2019) are increasingly being used to characterize the evolutionary dynamics of gene expression over time, for example, by looking for the signature of selection in the distribution of gene expression values at the tips (Dunn et al. 2013; Price et al. 2022; Brawand et al. 2011; Barua & Mikheyvey 2020; Rohlfs et al. 2014; Rohlfs & Nielsen 2015; Bedford & Hartl 2009). And accordingly, there have been a number of recent methodological developments, including computational platforms for simulating (Bastide et al. 2022) and analyzing (Bertram et al. 2022) phylogenetic comparative gene expression datasets. While this work is tremendously exciting, it is important to note that the reliability of the inferences from phylogenetic comparative methods hinge upon the performance of the phylogenetic model that is fit to the data (Garland et al. 1992; Price 1997; Boettiger et al. 2012; Pennell et al. 2015; Brown & Thomson 2018; Uyeda et al. 2021). There is a long tradition of using PCMs for modeling the evolution of morphological and ecological phenotypes but as comparative, multi-species gene expression datasets are starting to become more available the performance of the models in this new context is not well understood. And there are reasons to think that results from applying phylogenetic models to well-studied morphological phenotypes might not apply to gene expression data. First, evolutionary models of continuous traits were derived under the assumptions of quantitative genetics, where phenotypes are controlled by a large (effectively infinite) number of loci (Lande 1976; Turelli 1988; Felsenstein 1988; Lynch 1990; Hansen & Martins 1996; Pennell & Harmon 2013). We might expect the expression level of a given gene to behave less like an idealized polygenic trait owing to the outsized importance of the *cis-regulatory* region in determining the expression level (Dhar et al. 2021; Matharu & Ahituv 2020; Fuso et al. 2020; Romero et al. 2012). On the other hand, searches for eQTLs have turned up a large number of candidate loci potentially involved in the regulation of some genes (Rockman & Kruglyak 2006; GTEx Consortium 2020). Theoretical work has demonstrated that differences in the genetic architecture of traits influence the distribution of phenotypes among species (Schraiber & Landis 2015). Second, unlike traits such as height or mass, where the meaning of a measurement is straightforward, this is not the case for gene expression (Diaz et al. 2022); the number of mRNA transcripts is often normalized relative to the number of cells/transcripts/etc (Wagner et al. 2012). And it is not obvious how well different normalization measures match the distributional assumptions of phylogenetic models of trait evolution. And indeed, there is some empirical evidence to suspect that the assumptions of the independent contrasts method used by Dunn et al. 2018 in their reanalysis of pairwise comparisons were themselves problematic (Begum & Robinson-Rechavi 2021).

In a recent study, Chen et al. 2019 evaluated the fit of a set of alternative models to gene expression data. This set of models included Brownian motion (**BM**) (Felsenstein 1973) and varieties of the Ornstein-Uhlenbeck process (**OU**) (Hansen 1997; Butler & King 2004). Under BM a phenotypic trait with population mean $\underline{z}$ is expected to change over time period t according to a random walk such that:

$$\Delta\underline{z} = \sigma dW$$

where dW is a stochastic process drawn from a normal distribution with variance t and mean of 0, which is scaled by the parameter σ, such that σ^2 is defined as the evolutionary rate of the BM process. Over time, the variance between replicate lineages (i.e., two lineages that share a common ancestor and subsequently had independent evolutionary trajectories) of the phenotypic trait is expected to increase linearly such that:

$$Var(\underline{z}) = \sigma^2 t$$

The covariance between replicate lineages is proportional to the amount of shared evolutionary history. The OU process is an extension of the BM model where the mean change in phenotype over some period t is:

$$\Delta\underline{z} = -\alpha(\underline{z} - \theta) + \sigma dW$$

where ɑ is some pressure parameter keeping the trait value towards some optimal trait value θ with the same random walk σdW from BM contributing stochastic divergence. Chen et al. 2019 assessed the utility of phylogenetic models by comparing the relative fit of an alternative set of models. For this, they used the Akaike Information Criterion (**AIC**) (Akaike 1974). Comparing models with AIC is intended to find the model in a set that most closely approximates the generating model (Burnham & Anderson 2004) balancing accuracy with the additional prediction error that comes with adding free parameters. Alternative measures, such as likelihood ratio tests, BIC (Adkison et al. 1996), Bayes Factors (Kass & Raftery 1995), etc. differ in their details but are used for the same purpose.

However, model selection does not, however, indicate whether *any* of the compared models performs well (i.e., is *adequate*), in the sense that the distributional assumptions of the fitted model is consistent with the actual data. This is essential because even the best of a set of models may not adequately describe the structure of variation in the data and conclusions based on an inadequate model may not be reliable. Absolute model performance is typically assessed (when it is) with either parametric bootstrapping (Efron & Tibshirani 1993) when model parameters are estimated using maximum likelihood, or posterior predictive simulations (Rubin 1984; Gelman et al. 1996) when parameters are estimated using Bayesian inference. Essentially both parametric bootstrapping and posterior predictive simulations involve simulating new datasets given the model and fitted parameter values and assessing whether the observed data resembles the simulated datasets. If it does, then the model is considered to perform well for the observed dataset (for an overview of methods for assessing the performance of models in the context of evolutionary biology, see Brown & Thomson 2018)

Pennell et al. (2015) developed an approach, implemented in the R package Arbutus, designed to perform parametric bootstrapping or posterior predictive simulations for phylogenetic models of continuous trait evolution. In brief, the procedure is as follows: 1) a model of trait evolution is fit to a dataset; 2) the branch lengths of the original tree used in the analysis are “rescaled” such that if the model was a perfect fit to the data, the phylogenetic independent contrasts (**PICs**; Felsenstein 1985) computed on the tree would be independent and identically distributed and a standard normal distribution (i.e., $\sim Norm(0,1)$)); 3) the actual distribution of the contrasts are compared to the expected distribution using a variety of summary statistics. Each of these summary statistics measures deviations in the expected distribution of contrasts in unique ways (Pennell et al. 2015). *C.var* is the coefficient of variation of the absolute value of the PICs and is a measure of how well a model accounts for rate heterogeneity across a phylogeny. *D.cdf* is the *D* statistic from the Kolmolgorov-Smirnov test and measures deviations from the assumptions of normality for the contrasts such as in the case of rapid bursts of phenotypic character change. *S.asr* is the slope of a linear model between the absolute value of the contrasts and the inferred ancestral state of the nearest node to detect if magnitude of a trait is related to its evolutionary rate. *S.hgt* is the "node height test", which has been previously used to detect early bursts of phenotypic trait evolution such as in the case of an adaptive radiation (Freckleton & Harvey 2006; Slater & Pennell 2014). *S.var* is the slope of a linear regression between the absolute value of the contrasts against the expected variances of said contrasts and can be used to detect if the phylogenetic tree used in the fitted model has errors in the branch lengths. 4) If the observed summary statistic falls in either tail of the distribution of simulated summary statistics (e.g., $P<0.05$), the model can be considered inadequate.

Here we assess the performance of commonly used phylogenetic models of evolution for gene expression datasets from previously published studies that leveraged phylogenetic models across a variety of tissues, genes, and species. In addition to documenting the cases where these models fail to account for variation in comparative gene expression data, this study will be useful for identifying the reasons underlying this failure – and hopefully aid in the development of novel classes of phylogenetic models better suited to this type of data. We will focus on three core models, but also consider elaborations of these models later in the paper. These three core models are the aforementioned BM, OU, as well as Early Burst (**EB**) (Blomberg et al. 2003; Harmon et al. 2010). EB has not, to our knowledge, been applied to gene expression data but we included it because it makes a different set of distributional assumptions, such that it is a useful point of comparison. The EB process, often thought to characterize adaptive radiations (Schluter 2000; Harmon et al. 2010), is essentially the opposite of an OU model (Uyeda et al. 2015); the OU model leads to changes to the phenotypic variance being concentrated at the tips of the phylogeny whereas EB concentrates the variance near the root. Mathematically, the EB model is described by an exponential decrease in the rate of evolution through time where some trait mean $\underline{z}$ is determined by:

$$\Delta\underline{z}(t) = \sigma(t)dW$$

such that the diffusion (evolutionary rate) σ^2 as a function of time (t) is

$$\sigma^2(t) = \sigma_0^2 e^{rt}$$

where r is a positive parameter controlling the decrease in evolutionary rate.

Results

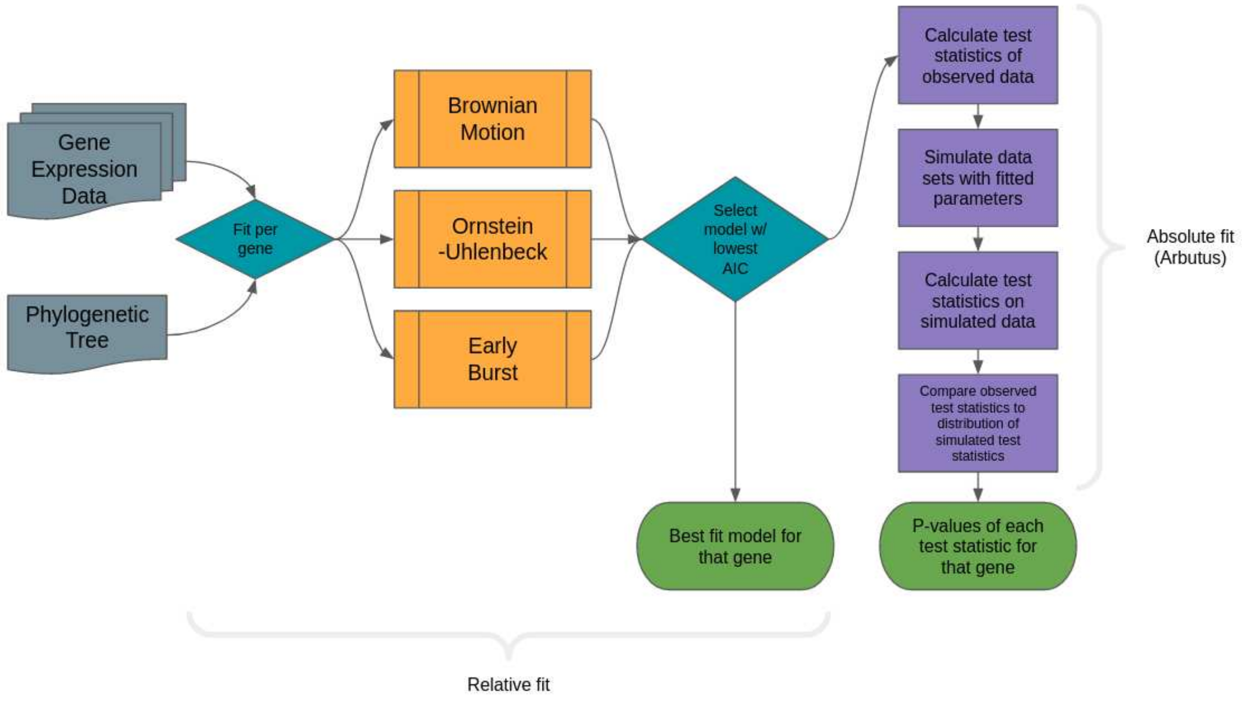
We aimed to explore model performance across a variety of different studies, including a range of taxa, tissues, and genes. To focus on relevant studies, we prioritized studies according to three criteria: first, that the originating study made use of at least one of the evolutionary models being assessed in this analysis and second, where the gene expression data and phylogenetic tree used in the study were readily available. The studies gathered in this process range in both number of genes and species analyzed as well as taxa included and tissues sampled (Table 1). Additionally, these studies made a variety of different claims regarding evolution of gene expression. For example, one study tested for the coevolution of proteins in fungi (Cope et al. 2020); another evaluated the ortholog conjecture by studying evolutionary rates following gene duplication (Kryuchkova-Mostacci & Robinson-Rechavi 2016). Rather than reevaluate the findings of any of the individual studies included in this analysis (Table 1), here we aim to find broad patterns in how well the distributional assumptions of widely used phylogenetic models conform to comparative gene expression data sets. By employing a wide range of studies, we hope to gain an understanding of how well these models perform in a plethora of different contexts.

Normalization has little effect on model adequacy

The data sets included in this analysis were normalized heterogeneously, with some count data being normalized as RPKM while others were normalized into TPM values. To ensure that the normalization method did not affect model adequacy, we re-analyzed genes from the CAVE FISH data set (Table 1). This data set was chosen because the authors provided raw RNA-Seq read counts for all the genes as well as reference transcriptomes. We found that model adequacy was nearly identical for both normalizations methods (Supplementary Figure 1).

Citation	Sequencing Platform	N. Genes	N. Taxa	Taxa Included	Organs	Study Designation
Fukushima & Pollock 2020	Multiple	1,377	21	Ensembl vertebrate species	Brain, Heart, Kidney, Liver, Ovary, Testis	AMALGAM
Stern & Crandall 2018	NextSeq 500	3560	14	Cave dwelling fish	Eye	CAVE FISH
El Taher et al. 2021	Illumina HiSeq 2500	32,596	73	Cichlids	Brain, Gill, Liver, Testis, Ovary, LPJ	CICHLIDS
Tobler et al. 2021	Illumina HiSeq 2500	16,740	20	Poecillidae fish	Gill	SULFIDE
Cope et al. 2020	Multiple	3,556	18	Fungus	NA	FUNGI
Catalán et al. 2019	Illumina HiSeq 2500	2,393	5	Heliconius butterflies	Brain	HELICONIUS
Kryuchkova-Mostacci & Robinson-Rechavi 2016	Multiple	8,333	9	Terrestrial animals	Varies	KMRR
Brawand et al. 2011	Illumina Genome Analyser IIx	5,320	10	Primates and outgroups	Brain, Cerebellum ,Heart, Kidney, Liver, Testis	MAMMALS
Barua & Mikheyev 2020	Multiple	11	52	Venomous snakes	Venom glands	VENOM

Table 1 Datasets included in this analysis. Data has to be making use of one of the evolutionary models, provide a phylogenetic tree, and have readily available gene expression data to be used in this analysis.


Gene Expression Data
Phylogenetic Tree
Fit per gene
Brownian Motion
Ornstein -Uhlenbeck
Early Burst
Select model w/ lowest AIC
Best fit model for that gene
Calculate test statistics of observed data
Simulate data sets with fitted parameters
Calculate test statistics on simulated data
Compare observed test statistics to distribution of simulated test statistics
P-values of each test statistic for that gene
Absolute fit (Arbutus)
Relative fit

224

225 **Figure 1 Workflow for determining relative and absolute fit of phylogenetic character models for gene expression data.** Data for each gene in a data set is analyzed by first fitting tested PCMs and then testing the best fit model for model adequacy using Arbutus. For data sets with available local gene trees, each gene is paired with its corresponding phylogenetic relationship.

OU models are the best supported model for the majority of genes

There are two levels of fit we considered for phylogenetic modeling: relative fit, – i.e., of the possible models for this set of data, which describes it the best — and absolute fit, – i.e., is the model describing the data well? For each of the studies listed in Table 1, we performed a series of analyses that can be summarized along those two tiers. First, we assessed the relative support for each of the three models on each of the genes in the data set to determine which of the three models best describes the evolution of that gene's expression (Figure 1) (See Methods for details). The best fit model for a gene was determined to be the model that minimized AIC. Second, we used Arbutus to measure the performance of the best-fit model for that gene's data (Figure 1). If multiple tissue types were included, model fit and performance was determined for each tissue type.

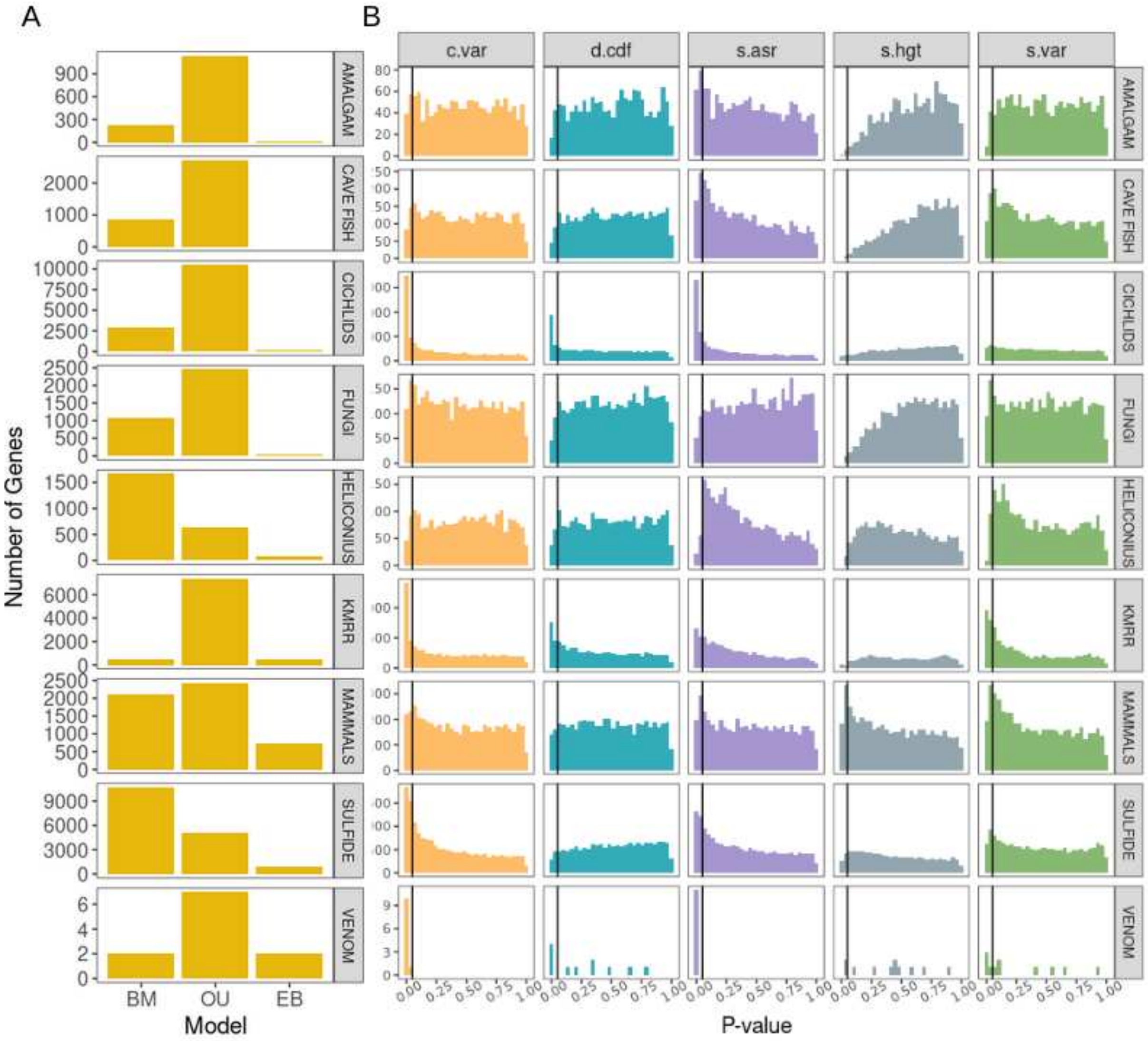


Figure 2 Relative (A) and absolute (B) fit of evolutionary models to the 9 gene expression data sets. Vertical black lines represent the significance cutoff of 0.05, with an expectation of 5% of genes being inadequate by chance. 59.8% of genes conform to the OU process. In terms of absolute performance, for 53% of genes the best fit model was adequate across all five test statistics. Model failures were primarily prevalent in *C.var* and *S.asr*.

Consistent with the work of Chen et al. (2019), we found that the OU model, commonly interpreted as an analog for stabilizing selection, was the best fit model for 59.8% of gene/tissue combinations, with noticeable exceptions in HELICONIUS and SULFIDE where the BM model was the best fit model for 66.5% and 63.8% of genes respectively (Figure 2). Notably, the HELICONIUS phylogeny is the smallest included in this study (Table 1) and we have low power to support more complex models. In a minority of cases (<15%), model failures were detected by the *D.cdf*, *S.hgt*, and *S.var* test statistics, with some notable exceptions in the KMRR data set (poor performance was detected for 21.1% of genes with *S.var*) and the CICHLIDS data set (poor performance was detected for 18.6% of genes with *D.cdf*). Starkly, every data set in this analysis except for FUNGI and HELICONIUS showed high concentrations of *P*-values below 0.05 for *C.var*, *S.asr*, or both (Figure 2). This is most extreme for the VENOM data set, where for every single, the model performed poorly in these aspects (Figure 2).

Models perform better when fit to species tree

Models fit to the KMRR data set showed poor performance across the board (Figure 2). One major difference between this study and data sets where the models also performed poorly (i.e., FUNGI, HELICONIUS, and AMALGAM), is the type of phylogenetic tree used. Unlike the other studies which each provided the species phylogeny they used for analysis, Kryuchkova-Mostacci & Robinson-Rechavi (2016) instead provided and used gene family phylogenies for each of the genes studied. Comparative analyses of "conventional" phenotypic traits, such as morphology, are typically conducted by using the species tree. However, if the genes underlying the phenotype are in regions of the genome that have different evolutionary histories than the species tree, estimates of phenotypic evolution may be biased. This is true of highly polygenic traits (Mendes et al. 2018; Hibbins et al. 2022) but appears especially problematic for traits that are underlain by a few genes (Hahn & Nakhleh 2016). So if the evolutionary models we used actually described evolution quite well, we would expect to see better model performance when using gene trees constructed from the regions of the genome that determine the expression of a particular gene. (On the other hand, phylogenetic error, particularly in the branch length estimation, may be particularly acute when estimating trees from small regions, which may introduce an additional set of problems.) Unfortunately, we do not know the loci responsible for variation in gene expression for most of the genes so a reasonable approximation would be to use the gene tree of the expressed gene itself as this should be closely linked to the promoter region, whose evolution will likely be important for the evolution of gene expression (Haberle & Stark 2018; Vaishnav et al. 2022).

Investigating this question comprehensively is beyond the scope of this paper as we do not have access to the original genomic sequence data for all of our datasets. But we did explore this by examining the FUNGI dataset using both the species tree and local gene trees for all the sampled genes (see Methods for how these trees were constructed). Substituting gene family phylogenies for the species phylogeny reduced the model performance as measured for all test statistics except for *D.cdf* . Two test statistics of note here would be *S.var* and *S.hgt*. The *S.var* statistic will indicate a model is inadequate when there are issues in branch length for the phylogeny used. The number of NA values for *S.hgt* was much higher when using species phylogenies, which could indicate low phylogenetic signal (see Münkemüller et al. 2012 for discussion of the measurement and interpretation of phylogenetic signal) when using this type of phylogenetic tree (Figure 3). This was confirmed to be the case with Blomberg's K test (Blomberg et al. 2003), where it shows lower K values for genes with NA values in *S.hgt* and thus, lower phylogenetic signal (Supplementary Figure 2). This higher incidence of NA values arises from the model fitting process. The summary statistic *S.hgt* is the slope of the relationship between the size of the contrasts and the height at which they occur. If an OU model is fit and the α parameter is very large, this essentially means that there is no phylogenetic signal. And in this case, the branch lengths leading to the tips of this transformed "unit tree" (see Pennell et al. 2015 for full mathematical details), will be very long. This means that there will be very little variance in node heights on the transformed trees and it will be therefore impossible to robustly estimate a slope; as such these cases are reported as NAs and excluded from subsequent analysis.

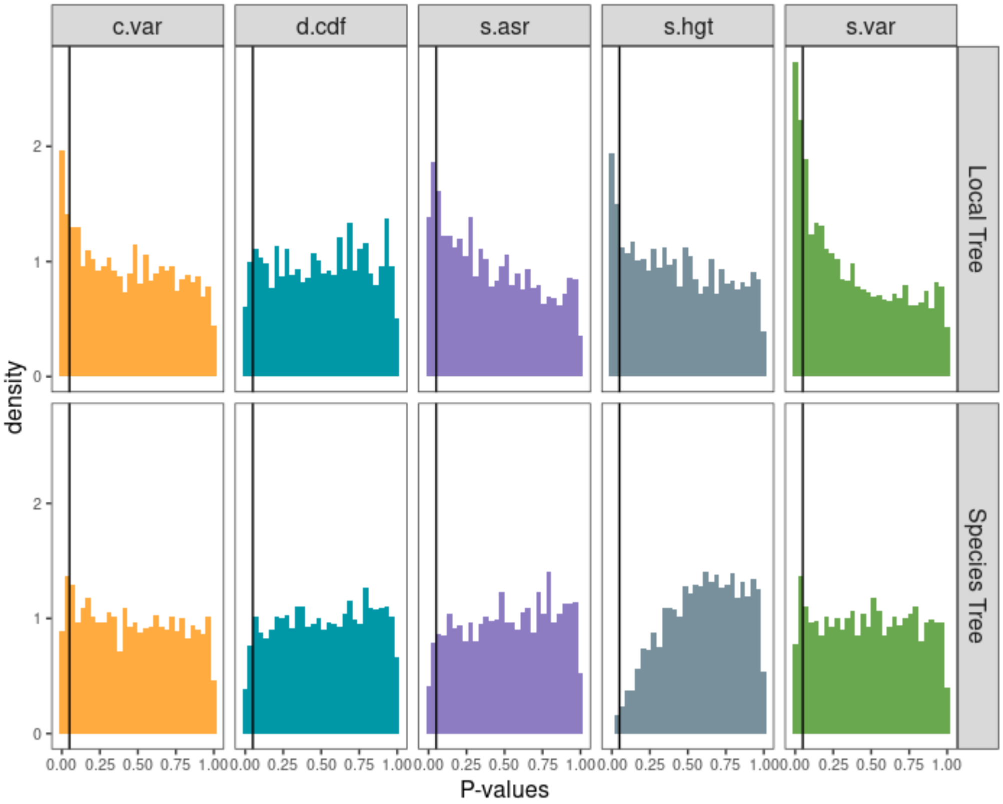


Figure 3 Analysis of generated local gene phylogenies. Test statistic *P-values* for best-fit models fit with the local gene phylogenies against those fit with the species phylogeny. Models showed poorer performance when they were fit to the gene trees versus the species trees, as measured by the summary statistics *C.var*, *S.asr*, *S.hgt*, and *S.var*.

Taken all together, it seems that for many genes, gene expression has higher phylogenetic signal when models are fit to the local gene trees but overall, models have better performance when they are fit to the species tree, primarily owing to the local trees have a higher frequency of violations detected by the *S.var* summary statistic, which we expect to be violated when there is a lot of branch length error (Pennell et al. 2015).

Discussion

In this study, we showed that model adequacy should be considered when applying PCMs to gene expression data to strengthen evolutionary inferences. The OU model is one of the most widely applied to make inferences about gene expression data (Bedford & Hartl 2009, Chen et al. 2019, Price et al. 2022). In this analysis we have shown that the a) OU is, by and large, the best of the 3 candidate models we tested (OU fits 59% of genes the best) and performs well for a majority (53%) of the gene/tissue combinations tested. This is an encouraging result as it lends further support to many of the conclusions that have been drawn from OU model fits to comparative gene expression datasets.

Similarly, to the angiosperm phenotypic data analyzed in Pennell et al. 2015, when the OU model performed poorly, deviations from the model expectations were primarily detected by *C.var*, *S.asr*, and *S.var*; which may be caused by statistical issues rather than biological processes. Thus, the OU model may simply be adequate primarily because it is the most flexible of the models tested in terms of phylogenetic and gene expression error (Pennell et al. 2015). In other words, we cannot determine from this analysis if OU is a good model because gene expression evolution is largely driven by stabilizing selection or due to the model's ability to “sop up” excess biological noise (Price et al. 2022; Cooper et al. 2016). In order to more accurately sift biologically relevant evolutionary claims from statistical artifacts, there are two straightforward (non-mutually exclusive) paths forward.

First, a number of recent analyses, including some of original publications from which our datasets were derived, used multi-rate or multi-optimum variants of the OU process (e.g., Chen et al. 2019, Tobler et al. 2021, Catalán et al. 2019, Brawand et al. 2011, Stern & Crandall 2018, Fukushima & Pollock 2020). This is important as a previous analysis by Chira and Thomas (2016) of morphological phenotypes found that failure to when the generating process was a multi-rate evolutionary model, fitting single-rate models to the data (as we have done here) would lead to poor model performance, as detected by the same summary statistics with which violations were commonly detected in our data – and that including multi-rate processes often led to better relative fit compared to single rate models and better model performance on absolute terms. While the Arbutus approach can be applied to a wide class of continuous trait models (see Pennell et al. 2015 for details), including both multi-rate and multi-optimum OU models, it was not clear how to apply this in a coherent and consistent way across the datasets and so we were unable to evaluate this here. This is because a key aspect of fitting multi-rate and multi-optimum models is deciding on the evolutionary regimes (i.e., the pattern of rate variation across lineages). In macroevolution, it is typical to either assign rate regimes to match a pre-specified biological hypothesis (O'Meara et al. 2006; Beaulieu et al. 2012) or to estimate the regimes alongside the parameters (Eastman et al. 2011, Thomas et al. 2012, Uyeda & Harmon 2014, Khabbazian et al. 2016). The former cannot be done in a standardized way across all of our datasets (the biological hypotheses are all distinct) and the latter is challenging to do when the number of lineages is relatively small (as is the case for our datasets). (It may be possible to combine data from different genes or tissues to estimate the evolutionary regimes (Bertram et al. 2022), but investigating this is beyond the scope of the present manuscript.)

Second, if one has multiple measurements for a given taxa (which unfortunately we did not for many of the datasets included here), one could use the Expression Variance Evolution model (Rohlfs et al. 2014, Rohlfs and Nielsen 2015) to jointly model the macroevolutionary dynamics and the processes that generate biological error within species. Given the limitations in many of the datasets, we were unable to do this. (The best we could do was to estimate a standard error for the estimates of the mean expression [see Methods for details] and include this as a fixed parameter when we fit the phylogenetic models.)

Another factor we thought might influence model performance was the type of phylogeny used in the study. Gene trees and species trees may have different topologies and different branch lengths and comparative analyses may show different results depending on the tree that is used (Hahn & Nakhleh 2016). Here we have shown that fitting models to the species tree rather than the gene tree (of the gene whose expression we are measuring) improves the performance of the phylogenetic models. Using local gene phylogenies increased model violations primarily as detected by *S.var*, indicating issues with branch lengths in gene level trees. This suggests that while gene phylogenies more accurately describe the relationship between the orthologs in a comparative gene expression study, the branch lengths of gene level phylogenies may be too error-prone to come to accurate conclusions. Indeed, Begum and Robinson-Rechavi (2021) found that biased and incorrect phylogenies can lead to erroneous conclusions. In lieu of this, we suggest that species phylogenetic trees should be used in PCMs until high accuracy gene trees are available. Furthermore, if gene expression at a focal gene is influenced by many different genes (i.e., *trans*-regulatory effects), it no longer makes conceptual sense to use the local gene tree.

An additional factor that will likely affect model performance is the size of the dataset (in terms of numbers of taxa). As gene expression data is still relatively expensive to collect (i.e., compared to many morphological traits), the size of many phylogenetic comparative studies of gene expression is relatively modest by modern macroevolutionary standards. As more and more taxa are included, the greater the chances that there will be substantial heterogeneity in the evolutionary process. It will also be the case that as datasets get larger, there will be more evidence to detect deviations from the assumptions of a model. Unfortunately, when assessing model performance it is rather difficult to disentangle these two factors and whether the distinction matters or not will depend on the research question (Pennell et al. 2015). Thus, we suspect that the reasonably good performance of relatively simple models may be due, at least in part, to the modestly sized datasets that we analyzed.

Conclusion

In this paper, we have conducted the first evaluation of absolute performance (in contrast to the relative fit) of phylogenetic models applied to gene expression data. The results are mixed. In a majority (61%) of the 155,679 gene-tissue combinations we analyzed, the best of the relatively simple models performed well. On the other hand, there were plenty of cases where we detected that the assumptions of phylogenetic models were severely different from reality. We did not try to replicate the exact methodology of the studies we pulled data from; naturally, these are all asking different questions and have nuanced context and approaches and it was therefore impossible to do a formal meta-analyses. Nonetheless, for cases where there were substantial misalignment between model and data, it would certainly be worth revisiting whether any specific biological conclusion hinged on the parameters estimated from these models – if, for instance, an OU model is used to detect stabilizing selection but an OU model performs poorly for a specific dataset of interest, this would imply that the inference of stabilizing selection (or not) may be misleading (Pennell et al. 2015, Price et al. 2022). A secondary aim of our paper is to promote the evaluation of model performance as a critical part of any data analysis pipeline (Brown &Thomson 2018). The tool we used here, Arbutus, can be adapted to evaluate performance for a wide array of phylogenetic models for continuous traits (Pennell et al. 2015). Assessing the absolute performance of a phylogenetic model in a comparative gene expression study can provide more confidence in the results of the analyses (if the models used broadly perform well) or suggest new models that should be considered (if they do not). We, like many others, are excited by the fact that comparative gene expression studies are becoming increasingly phylogenetic – there are many exciting evolutionary questions that we may finally be able to address (Price et al. 2022). We hope this contributions aids in this work but helping researchers ensure that their inferences regarding these questions are on a sure footing.

Methods

Analysis of Model Fit

We log-transformed normalized gene expression values for all data sets (see below for details on normalization) before we evaluated model fit and adequacy to facilitate cross-species comparisons. For every gene-tissue combination, we used the 'fitContinuous()' function in geiger (Pennell et al. 2014) to fit BM, OU, and EB to the comparative gene expression dataset. When a species tip was missing data for a gene, that tip was excised before performing fitting and adequacy measurement. If data sets included multiple samples per species, the mean expression was used and an error term equivalent to the standard error of the gene expression data for that species was used for that gene. Relative model fit was assessed on a per-gene basis, with each gene being assigned one model with the best fit; i.e., the model with the lowest AIC score as calculated by the model-fitting process. We then plotted best-fit models using the ggplot R package (Wickham 2016). Model adequacy was calculated using best-fit model parameters calculated in the previous step using the Arbutus R package (Pennell et al. 2015).

Evaluating the effect of standardization

To measure the effect of normalization type on model adequacy, we compared model adequacy between RPKM and TPM values for the CAVE FISH data set (Table 1). We quality trimmed these reads using Trimmomatic (Bolger et al. 2014) and performed alignment and quantification using the Trinity pipeline (Grabherr et al. 2011), producing both RPKM and TPM values for all genes included. To compare adequacy we then performed adequacy analysis as explained above for both normalization methods.

Local Gene Phylogeny Construction

We generated gene family phylogenies for the Cope et al. 2020 data set using protein sequences downloaded from the Ensembl database (Howe et al. 2021) via the biomaRt package (Durinck et al. 2022). We then aligned downloaded sequences using MAFFT (Katoh & Standley 2013) and assembled them into phylogenetic trees using FastTree (Price et al. 2010), which uses a minimum evolution model to build trees. We fit chronograms to gene trees to make them ultrametric using penalized likelihood as implemented in the ape package with the chronos function (Paradis & Schliep 2019). We implemented this in a single Snakemake (Mölder et al. 2021) pipeline.

Testing for Phylogenetic Signal

Phylogenetic signal was compared for genes with NA values in the *S.hgt* metric against genes with real numerical values using the phytools R package (Revell 2012). Results were plotted for both the K-statistic (Blomberg et al. 2003).

Data and Code Availability

All R scripts, pipelines, and data used in this analysis can be found in or redirected from the following GitHub repository: https://github.com/fieldima/adequacy_of_PCMs.

Acknowledgements

We thank Paul Pavlidis, Keith Adams, Alex Cope, and members of the Pennell and Pavlidis labs for comments on the work and the manuscript. Casey Dunn and Felipe

Zapata provided additional insights into the problems addressed here. MP was funded by a NSERC Discovery Grant and from startup funds from the University of Southern California.

References

Adkison MD, Peterman RM, Lapointe MF, Gillis DM, Korman J. 1996. Alternative models of climatic effects on sockeye salmon, Oncorhynchus nerka, productivity in Bristol Bay, Alaska, and the Fraser River, British Columbia. Fisheries Oceanogr. 5:137–152. doi: 10.1111/j.1365-2419.1996.tb00113.x.

Akaike H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 19:716–723. doi: 10.1109/TAC.1974.1100705.

Barua A, Mikheyev AS. 2020. Toxin expression in snake venom evolves rapidly with constant shifts in evolutionary rates. Proceedings of the Royal Society B: Biological Sciences. 287:20200613. doi: 10.1098/rspb.2020.0613.

Bastide P, Soneson C, Stern DB, Lespinet O, Gallopin M. 2022. A Phylogenetic Framework to Simulate Synthetic Inter-species RNA-Seq Data Battistuzzi, FU, editor. Molecular Biology and Evolution. msac269. doi: 10.1093/molbev/msac269.

Beaulieu JM, Jhwueng D-C, Boettiger C, O'Meara BC. 2012. Modeling Stabilizing Selection: Expanding The Ornstein-Uhlenbeck Model of Adaptive Evolution. Evolution. 66:2369–2383. doi: 10.1111/j.1558-5646.2012.01619.x.

Bedford T, Hartl DL. 2009. Optimization of gene expression by natural selection. Proc. Natl. Acad. Sci. U.S.A. 106:1133–1138. doi: 10.1073/pnas.0812009106.

Begum T, Robinson-Rechavi M. 2021. Special Care Is Needed in Applying Phylogenetic Comparative Methods to Gene Trees with Speciation and Duplication Nodes Satta, Y, editor. Molecular Biology and Evolution. 38:1614–1626. doi: 10.1093/molbev/msaa288.

Bertram J et al. 2022. CAGEE: computational analysis of gene expression evolution. doi: 10.1101/2022.11.18.517074.

Blomberg SP, Garland T, Ives AR. 2003. Testing For Phylogenetic Signal in Comparative Data: Behavioral Traits Are More Labile. Evolution. 57:717–745. doi: 10.1111/j.0014-3820.2003.tb00285.x.

Boettiger C, Coop G, Ralph P. 2012. Is Your Phylogeny Informative? Measuring The Power Of Comparative Methods: Is Your Phylogeny Informative? Evolution. 66:2240–2251. doi: 10.1111/j.1558-5646.2011.01574.x.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 30:2114–2120. doi: 10.1093/bioinformatics/btu170.

Brawand D et al. 2011. The evolution of gene expression levels in mammalian organs. Nature. 478:343–348. doi: 10.1038/nature10532.

Brown JM, Thomson RC. 2018. Evaluating Model Performance in Evolutionary Biology. Annu. Rev. Ecol. Evol. Syst. 49:95–114. doi: 10.1146/annurev-ecolsys-110617-062249.

Burnham KP, Anderson DR. 2004. Multimodel Inference: Understanding AIC and BIC in Model Selection. Sociological Methods & Research. 33:261–304. doi: 10.1177/0049124104268644.

Butler MA, King AA. 2004. Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution. The American Naturalist. 164:683–695. doi: 10.1086/426002.

Cain CE, Blekhman R, Marioni JC, Gilad Y. 2011. Gene expression differences among primates are associated with changes in a histone epigenetic modification. Genetics. 187:1225–1234. doi: 10.1534/genetics.110.126177.

Catalán A, Briscoe AD, Höhna S. 2019. Drift and Directional Selection Are the Evolutionary Forces Driving Gene Expression Divergence in Eye and Brain Tissue of Heliconius Butterflies. Genetics. 213:581–594. doi: 10.1534/genetics.119.302493.

Chen J et al. 2019. A quantitative framework for characterizing the evolutionary history of mammalian gene expression. Genome Res. 29:53–63. doi: 10.1101/gr.237636.118.

Chen D et al. 2021. Single cell atlas for 11 non-model mammals, reptiles and birds. Nat Commun. 12:7083. doi: 10.1038/s41467-021-27162-2.

Chira AM, Thomas GH. 2016. The impact of rate heterogeneity on inference of phylogenetic models of trait evolution. J. Evol. Biol. 29:2502–2518. doi: 10.1111/jeb.12979.

Cooper N, Thomas GH, Venditti C, Meade A, Freckleton RP. 2016. A cautionary note on the use of Ornstein Uhlenbeck models in macroevolutionary studies. Biol. J. Linn. Soc. 118:64–77. doi: 10.1111/bij.12701.

Cope AL, O'Meara BC, Gilchrist MA. 2020. Gene expression of functionally-related genes coevolves across fungal species: detecting coevolution of gene expression using phylogenetic comparative methods. BMC Genomics. 21:370. doi: 10.1186/s12864-020-6761-3.

Dhar GA, Saha S, Mitra P, Nag Chaudhuri R. 2021. DNA methylation and regulation of gene expression: Guardian of our health. Nucleus. 64:259–270. doi: 10.1007/s13237-021-00367-y.

Diaz R, Wang Z, Townsend JP. 2023. Chapter 5 - Measurement and meaning in gene expression evolution. In: Transcriptome Profiling. Ajmal Ali, M & Lee, J, editors. Academic Press pp. 111–129. doi: 10.1016/B978-0-323-91810-7.00008-X.

Dunn CW, Luo X, Wu Z. 2013. Phylogenetic Analysis of Gene Expression. Integrative and Comparative Biology. 53:847–856. doi: 10.1093/icb/ict068.

Dunn CW, Zapata F, Munro C, Siebert S, Hejnol A. 2018. Pairwise comparisons across species are problematic when analyzing functional genomic data. Proc. Natl. Acad. Sci. U.S.A. 115. doi: 10.1073/pnas.1707515115.

Durinck S et al. 2022. biomaRt: Interface to BioMart databases (i.e. Ensembl). Bioconductor version: Release (3.15) doi: 10.18129/B9.bioc.biomaRt.

Eastman JM, Alfaro ME, Joyce P, Hipp AL, Harmon LJ. 2011. A novel comparative method for identifying shifts in the rate of character evolution on trees. Evolution. 65:3578–3589. doi: 10.1111/j.1558-5646.2011.01401.x.

Efron B, Tibshirani R. 1993. An Introduction to the Bootstrap. Chapman & Hall: New York.

El Taher A et al. 2021. Gene expression dynamics during rapid organismal diversification in African cichlid fishes. Nat Ecol Evol. 5:243–250. doi: 10.1038/s41559-020-01354-3.

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

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1762641/ (Accessed December 5, 2022).

Felsenstein J. 1985. Phylogenies and the Comparative Method. The American Naturalist. 125:1–15. doi: 10.1086/284325.

Felsenstein J. 1988. Phylogenies and Quantitative Characters. Annu. Rev. Ecol. Syst. 19:445–471. doi: 10.1146/annurev.es.19.110188.002305.

Freckleton RP, Harvey PH. 2006. Detecting Non-Brownian Trait Evolution in Adaptive Radiations. PLOS Biology. 4:e373. doi: 10.1371/journal.pbio.0040373.

Fukushima K, Pollock DD. 2020. Amalgamated cross-species transcriptomes reveal organ-specific propensity in gene expression evolution. Nat Commun. 11:4459. doi: 10.1038/s41467-020-18090-8.

Fuso A, Raia T, Orticello M, Lucarelli M. 2020. The complex interplay between DNA methylation and miRNAs in gene expression regulation. Biochimie. 173:12–16. doi: 10.1016/j.biochi.2020.02.006.

Garamszegi LZ, ed. 2014. Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology: Concepts and Practice. 1st ed. 2014. Springer Berlin Heidelberg : Imprint: Springer: Berlin, Heidelberg.

Garland T, Harvey PH, Ives AR. 1992. Procedures for the Analysis of Comparative Data Using Phylogenetically Independent Contrasts. Systematic Biology. 41:18–32. doi: 10.1093/sysbio/41.1.18.

Gelman A, Meng XL, Stern H. 1996. Posterior Predictive Assessment of Model Fitness Via Realized Discrepancies. Statistica Sinica. 6:733–760. https://www.jstor.org/stable/24306036 (Accessed October 27, 2022).

Grabherr MG et al. 2011. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nature biotechnology. 29:644. doi: 10.1038/nbt.1883.
GTEx Consortium 2020. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 369:1318–1330. doi: 10.1126/science.aaz1776.

Haberle V, Stark A. 2018. Eukaryotic core promoters and the functional basis of transcription initiation. Nat Rev Mol Cell Biol. 19:621–637. doi: 10.1038/s41580-018-0028-8.

Hahn MW, Nakhleh L. 2016. Irrational exuberance for resolved species trees: COMMENTARY. Evolution. 70:7–17. doi: 10.1111/evo.12832.

Hansen TF, Martins EP. 1996. Translating Between Microevolutionary Process And Macroevolutionary Patterns: The Correlation Structure Of Interspecific Data. Evolution. 50:1404–1417. doi: 10.1111/j.1558-5646.1996.tb03914.x.

Hansen TF. 1997. Stabilizing Selection And The Comparative Analysis Of Adaptation. Evolution. 51:1341–1351. doi: 10.1111/j.1558-5646.1997.tb01457.x.

Harmon LJ et al. 2010. Early Bursts Of Body Size And Shape Evolution Are Rare In Comparative Data: Early Bursts Of Evolution Are Rare. Evolution. no-no. doi: 10.1111/j.1558-5646.2010.01025.x.

Harmon L. 2019. Phylogenetic Comparative Methods. Independent.

Hernando-Herraez et al. 2015. The interplay between DNA methylation and sequence divergence in recent human evolution. Nucleic Acids Res. 43:8204–8214. doi: 10.1093/nar/gkv693.

Hibbins MS, Breithaupt LC, Hahn MW. 2022. Phylogenomic comparative methods: accurate evolutionary inferences in the presence of gene tree discordance. doi: 10.1101/2022.11.14.516436.

Hill MS, Vande Zande P, Wittkopp PJ. 2021. Molecular and evolutionary processes generating variation in gene expression. Nat Rev Genet. 22:203–215. doi: 10.1038/s41576-020-00304-w.

Howe KL et al. 2021. Ensembl 2021. Nucleic Acids Research. 49:D884–D891. doi: 10.1093/nar/gkaa942.

Kass RE, Raftery AE. 1995. Bayes Factors. Journal of the American Statistical Association. 90:773–795. doi: 10.1080/01621459.1995.10476572.

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution. 30:772–780. doi: 10.1093/molbev/mst010.

Kellis M et al. 2014. Defining functional DNA elements in the human genome. Proc. Natl. Acad. Sci. U.S.A. 111:6131–6138. doi: 10.1073/pnas.1318948111.

Khabbazian M, Kriebel R, Rohe K, Ané C. 2016. Fast and accurate detection of evolutionary shifts in Ornstein–Uhlenbeck models Hansen, T, editor. Methods Ecol Evol. 7:811–824. doi: 10.1111/2041-210X.12534.

King M-C, Wilson AC. 1975. Evolution at Two Levels in Humans and Chimpanzees: Their macromolecules are so alike that regulatory mutations may account for their biological differences. Science. 188:107–116. doi: 10.1126/science.1090005.

Kryuchkova-Mostacci N, Robinson-Rechavi M. 2016. Tissue-Specificity of Gene Expression Diverges Slowly between Orthologs, and Rapidly between Paralogs. PLOS Computational Biology. 12:e1005274. doi: 10.1371/journal.pcbi.1005274.

Lande R. 1976. Natural Selection and Random Genetic Drift in Phenotypic Evolution. Evolution. 30:314. doi: 10.2307/2407703.

Lynch M. 1990. The similarity index and DNA fingerprinting. Mol Biol Evol. 7:478–484. doi: 10.1093/oxfordjournals.molbev.a040620.

Matharu N, Ahituv N. 2020. Modulating gene regulation to treat genetic disorders. Nat Rev Drug Discov. 19:757–775. doi: 10.1038/s41573-020-0083-7.

Mendes FK, Fuentes-González JA, Schraiber JG, Hahn MW. 2018. A multispecies coalescent model for quantitative traits Wittkopp, PJ, Rokas, A, & Pennell, M, editors. eLife. 7:e36482. doi: 10.7554/eLife.36482.

Mölder F et al. 2021. Sustainable data analysis with Snakemake. F1000Research https://f1000research.com/articles/10-33 (Accessed August 18, 2022).

Münkemüller T et al. 2012. How to measure and test phylogenetic signal: How to measure and test phylogenetic signal. Methods in Ecology and Evolution. 3:743–756. doi: 10.1111/j.2041-210X.2012.00196.x.

O'Meara BC, Ané C, Sanderson MJ, Wainwright PC. 2006. Testing For Different Rates of Continuous Trait Evolution Using Likelihood. Evolution. 60:922–933. doi: 10.1111/j.0014-3820.2006.tb01171.x.

Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R Schwartz, R, editor. Bioinformatics. 35:526–528. doi: 10.1093/bioinformatics/bty633.

Pennell MW, Harmon LJ. 2013. An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology: Integrative comparative methods. Ann. N.Y. Acad. Sci. 1289:90–105. doi: 10.1111/nyas.12157.

Pennell, M.W., J.M. Eastman, G.J. Slater, J.W. Brown, J.C. Uyeda, R.G. FitzJohn, M.E. Alfaro, and L.J. Harmon. 2014. geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. Bioinformatics 30:2216-2218.

Pennell MW, FitzJohn RG, Cornwell WK, Harmon LJ. 2015. Model Adequacy and the Macroevolution of Angiosperm Functional Traits. The American Naturalist. 186:E33–E50. doi: 10.1086/682022.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLOS ONE. 5:e9490. doi: 10.1371/journal.pone.0009490.

Price PD et al. 2022. Detecting signatures of selection on gene expression. Nat Ecol Evol. 6:1035–1045. doi: 10.1038/s41559-022-01761-8.

Price T. 1997. Correlated evolution and independent contrasts. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences. 352:519–529. doi: 10.1098/rstb.1997.0036.

Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things): phytools: R package. Methods in Ecology and Evolution. 3:217–223. doi: 10.1111/j.2041-210X.2011.00169.x.

Rockman MV, Kruglyak L. 2006. Genetics of global gene expression. Nat Rev Genet. 7:862–872. doi: 10.1038/nrg1964.

Rohlfs RV, Harrigan P, Nielsen R. 2014. Modeling Gene Expression Evolution with an Extended Ornstein–Uhlenbeck Process Accounting for Within-Species Variation. Molecular Biology and Evolution. 31:201–211. doi: 10.1093/molbev/mst190.

Rohlfs RV, Nielsen R. 2015. Phylogenetic ANOVA: The Expression Variance and Evolution Model for Quantitative Trait Evolution. Syst Biol. 64:695–708. doi: 10.1093/sysbio/syv042.

Romero IG, Ruvinsky I, Gilad Y. 2012. Comparative studies of gene expression and the evolution of gene regulation. Nat Rev Genet. 13:505–516. doi: 10.1038/nrg3229.

Rubin DB. 1984. Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. The Annals of Statistics. 12:1151–1172. doi: 10.1214/aos/1176346785.

Schluter D. 2000. The ecology of adaptive radiation. Oxford University Press: Oxford.

Schraiber JG, Landis MJ. 2015. Sensitivity of quantitative traits to mutational effects and number of loci. Theoretical Population Biology. 102:85–93. doi: 10.1016/j.tpb.2015.03.005.

Slater GJ, Pennell MW. 2014. Robust Regression and Posterior Predictive Simulation Increase Power to Detect Early Bursts of Trait Evolution. Systematic Biology. 63:293–308. doi: 10.1093/sysbio/syt066.

Stern DB, Crandall KA. 2018. The Evolution of Gene Expression Underlying Vision Loss in Cave Animals. Mol Biol Evol. 35:2005–2014. doi: 10.1093/molbev/msy106.

Thomas GH, Freckleton RP. 2012. MOTMOT: models of trait macroevolution on trees: MOTMOT. Methods in Ecology and Evolution. 3:145–151. doi: 10.1111/j.2041-210X.2011.00132.x.

Turelli M. 1988. Phenotypic Evolution, Constant Covariances, and the Maintenance of Additive Variance. Evolution. 42:1342. doi: 10.2307/2409017.

Tobler M, Greenway R, Kelley JL. 2021. Ecology drives the degree of convergence in the gene expression of extremophile fishes. bioRxiv https://www.biorxiv.org/content/10.1101/2021.12.13.472416v1 (Accessed May 29, 2022).

Uyeda JC, Harmon LJ. 2014. A Novel Bayesian Method for Inferring and Interpreting the Dynamics of Adaptive Landscapes from Phylogenetic Comparative Data. Systematic Biology. 63:902–918. doi: 10.1093/sysbio/syu057.

Uyeda JC, Caetano DS, Pennell MW. 2015. Comparative Analysis of Principal Components Can be Misleading. Systematic Biology. 64:677–689. doi: 10.1093/sysbio/syv019.

Uyeda JC, Zenil-Ferguson R, Pennell MW. 2018. Rethinking phylogenetic comparative methods. Syst Biol. 67:1091–1109. doi: 10.1093/sysbio/syy031.

Uyeda JC, Bone N, McHugh S, Rolland J, Pennell MW. 2021. How should functional relationships be evaluated using phylogenetic comparative methods? A case study using metabolic rate and body temperature. Evolution. 75:1097–1105. doi: 10.1111/evo.14213.

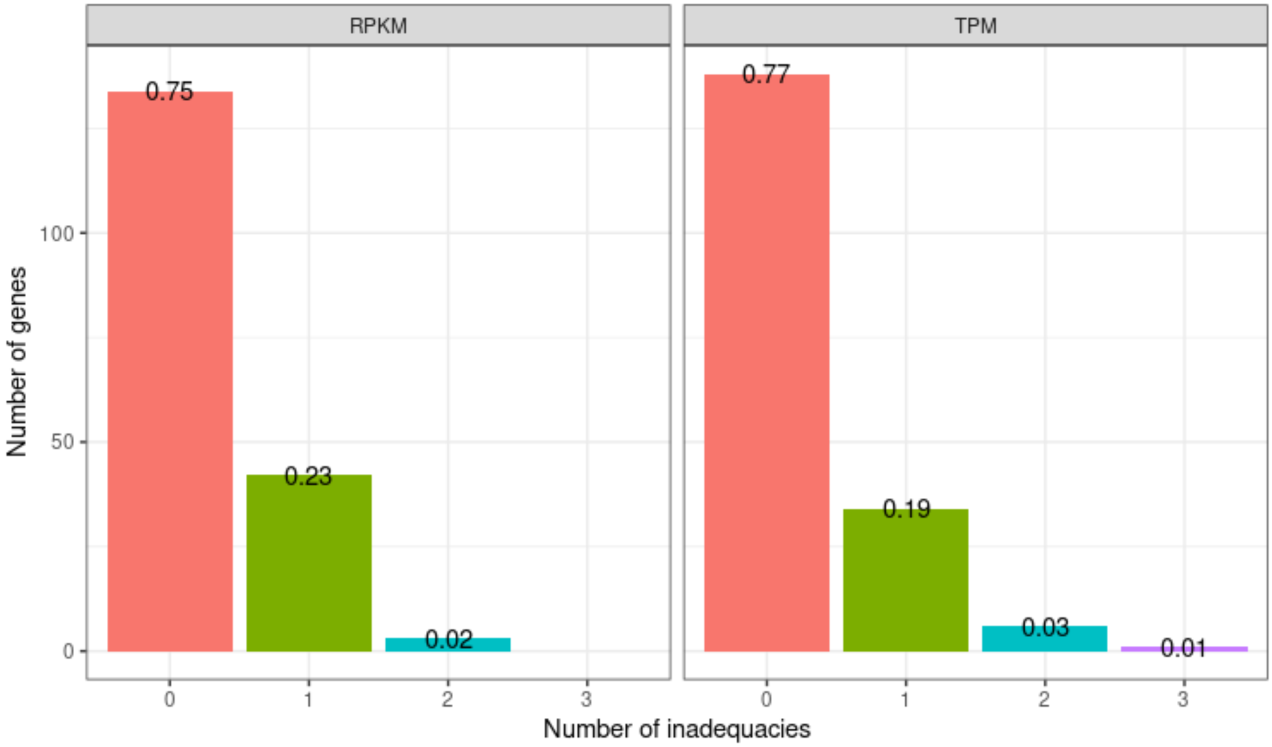
Vaishnav ED et al. 2022. The evolution, evolvability and engineering of gene regulatory DNA. Nature. 603:455–463. doi: 10.1038/s41586-022-04506-6.

Villar D et al. 2015. Enhancer Evolution across 20 Mammalian Species. Cell. 160:554–566. doi: 10.1016/j.cell.2015.01.006.

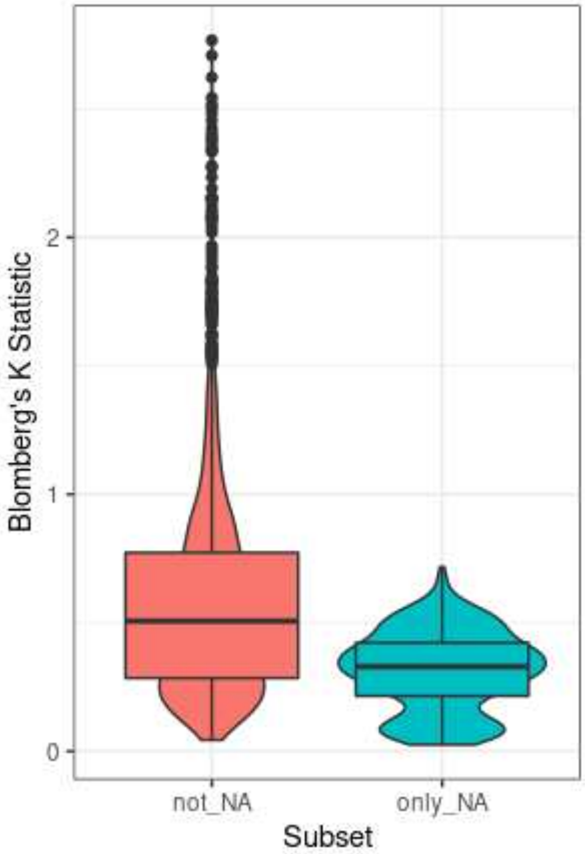
Wagner GP, Kin K, Lynch VJ. 2012. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci. 131:281–285. doi: 10.1007/s12064-012-0162-3.

Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis. 2nd ed. 2016. Springer International Publishing : Imprint: Springer: Cham.

Supplementary Figures



Supplementary Figure 1 Proportion of inadequate genes for RPKM (left) and TPM (right) normalized reads. Raw reads from the CAVE FISH data set were normalized into RPKM and TPM values and then the best fit model was analyzed for model adequacy via ARBUTUS. The proportion of genes with zero, one, or two inadequacies was nearly identical between both modes of normalization.



Supplementary Figure 2 Blomberg's K statistic for genes with NA values (left) and non-NA values (right) in the *S.hgt* test statistic. NA genes have a lower K statistic on average than non-NA genes. NA genes have lower phylogenetic signal.