Pairwise comparisons are problematic when analyzing functional genomic data across species

Casey W. Dunn[1]*, Felipe Zapata[1,2], Catriona Munro[1], Stefan Siebert[1,3], Andreas Hejnol[4]

[1] Department of Ecology and Evolutionary Biology, Brown University, Providence, RI, USA

[2] Current address: Department of Ecology and Evolutionary Biology, University of California Los Angeles, Los Angeles, CA, USA

[3] Current address: Department of Molecular & Cellular Biology, University of California at Davis, Davis, CA, USA

[4] Sars International Centre for Marine Molecular Biology, University of Bergen, Bergen, Norway

* Corresponding author, casey_dunn@brown.edu

Abstract

There is now considerable interest in comparing functional genomics data across species. One goal of this work is to provide an integrated understanding of genome and phenotype evolution. Most studies of this type have relied on multiple pairwise comparisons of functional genomic data between species, an approach that does not incorporate information about the evolutionary relationships among species. The statistical problems that arise from not considering these relationships can lead pairwise approaches to the wrong conclusions, and are a missed opportunity to learn about biology that can only be understood in an explicit phylogenetic context. Here we examine two recently published studies that compare gene expression across species with pairwise methods. We find problems with both that call their conclusions into question. One study interpreted higher expression correlation between orthologs than paralogs as evidence for the ortholog conjecture, *i.e.* the hypothesis that gene function evolution is more rapid after duplication than speciation. Instead, we find that this pattern is due to the structure of the gene phylogenies, not different rates of expression evolution, and is the expected pattern when rates are the same following duplication and speciation. The second study interpreted pairwise comparisons of embryonic gene expression across distantly related animals as evidence for a distinct evolutionary process that gave rise to animal phyla. Instead, we find that the pattern they identified is due to unique features of a single species that impact multiple pairwise comparisons that include that species. In each study, distinct patterns of pairwise similarity among species were interpreted as evidence of particular evolutionary processes, but instead reflect the structure of phylogenies. These reanalyses concretely demonstrate the inadequacy of pairwise comparisons for analyzing functional genomic data across species, and indicate that it will be critical to adopt phylogenetic comparative methods in future functional genomics work. Fortunately, phylogenetic comparative biology is also a rapidly advancing field with many methods that can be directly applied to functional genomic data.

Introduction

The focus of genomics research has quickly shifted from describing genome sequences to functional genomics, the study of how genomes "work" using tools that measure functional attributes such as expression, chromatin state, and transcription initiation. Functional genomics, in turn, is now becoming more comparative– there is great interest in understanding how functional genomic variation across species gives rise to a diversity of development, morphology, physiology, and other phenotypes[1]. These analyses are also critical to transferring functional insight across species, and will grow in importance in coming years.

A rich theoretical and statistical methodology of phylogenetic comparative methods has been developed over the last three decades to address the challenges and opportunities of trait comparisons across species[2–7]. A central challenge is the dependence of observations across species due to the evolutionary history of species – more closely related species share many traits that evolved once in a common ancestor. This violates

the fundamental assumption of observation independence in standard statistical methods. Phylogenetic comparative methods address this dependence. They have largely been applied to morphological and ecological traits, but are just as relevant to functional genomics. Even so, most comparative functional genomics studies have abstained from phylogenetic approaches and instead rely on multiple pairwise comparisons across species (Figure 1a). This leaves comparative functional genomics studies susceptible to statistical problems and is a missed opportunity to ask questions that are only accessible in an explicit phylogenetic context.

Phylogenetic comparative methods account for evolutionary history and explicitly model trait change along the branches of evolutionary trees (*e.g.*, Figure 1b). The value of these methods relative to to pairwise comparisons has been repeatedly shown in analyses of other types of character data[8–10]. The application of these methods is illustrated by a study of leaf functional attributes for about 100 plant species[11]. If phylogenetic relationships are not considered, analyses indicate a negative correlation of leaf lifespan with leaf size. With the application of phylogenetic comparative methods, however, this correlation disappears because most of the observed variance is due to differences between conifers and flowering plants. Phylogenetic comparative methods capture this shift as change along a single branch deep in the tree, revealing that there is not a tendency of correlated change between these traits across the phylogeny. One reason that comparative functional genomic sudies have not embraced phylogenetic approaches is that it has not yet been concretely demonstrated that pairwise and phylogenetic comparative methods can lead to different results when considering functional genomics data. Here we examine this issue by re-evaluating the pairwise comparisons in two recent studies that compared gene expression across species.

The first study, Kryuchkova-Mostacci and Robinson-Rechavi (KMRR) 2016[12], analyzed multiple vertebrate expression datasets to test the ortholog conjecture - the hypothesis that orthologs tend to have more conserved attributes (specificity of expression across organs in this case) than do paralogs[13]. Using pairwise comparisons (Figure 1a), they found lower expression correlation between paralogs than between orthologs and interpreted this as strong support for the ortholog conjecture.

The second comparative functional genomics study we evaluate here is Levin *et al.*[15]. This study analyzed gene expression through the course of embryonic development for ten animal species, each from a different phylum. Using pairwise comparisons, they found there is more evolutionary variance in gene expression at a mid phase of development than there is at early and late phases. They suggest that this supports an "inverse hourglass" model for the evolution of gene expression, in contrast with the "hourglass" model previously proposed for closely related species[16]. Furthermore, they suggested that this provides biological justification for the concept of phyla. We previously described concerns with the interpretations of this result[17]. Here we address the analyses themselves by examining the structure of the pairwise comparisons.

Results and Discussion

KMRR Reanalysis

Original pairwise test of the ortholog conjecture

The KMRR study[12] sought to test the ortholog conjecture. The ortholog conjecture[13] is the proposition that orthologs (genes that diverged from each other due to a speciation event) have more similar attributes than do paralogs (genes that diverged from each other due to a gene duplication event). The ortholog conjecture has important biological and technical implications. It shapes our understanding of the functional diversity of gene families. It is also used to relate findings from well-studied genes to related genes that have not been investigated in detail. It can be applied to any trait of genes, from gene sequence to biochemical properties to expression. While the ortholog conjecture describes a specific pattern of functional diversity across genes, it is also articulated as a hypothesis about the process of evolution– that there is greater evolutionary change in gene attributes following a duplication event than a speciation event.

Despite its importance, there have been relatively few tests of the ortholog conjecture. Previous work has shown that ontology annotations are not sufficient to test the ortholog conjecture[18,19]. Analyses of domain structure were consistent with the ortholog conjecture[20]. There have been few tests of the ortholog conjecture with regards to gene expression[18], and KMRR is the most thorough such expression study to date.

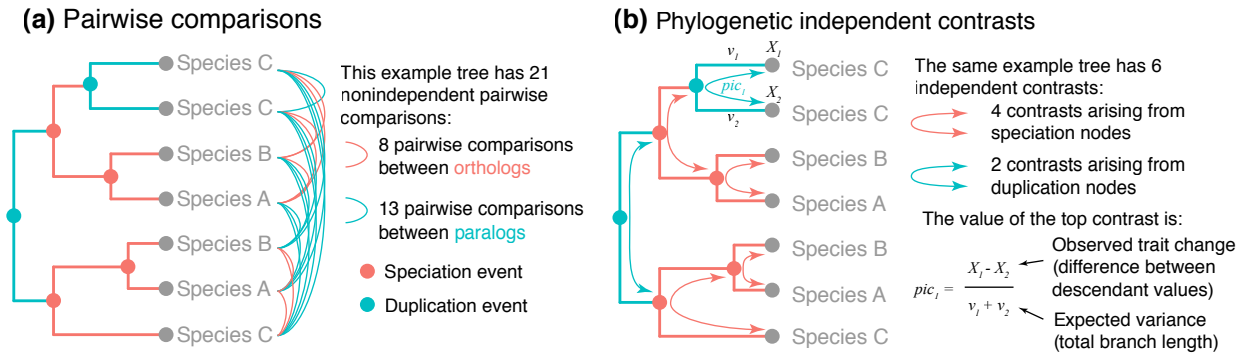

Figure 1: Pairwise and phylogenetic comparative approaches, illustrated on an example gene tree with multiple genes per species. The internal nodes of the tree are speciation and gene duplication events. (a) Many comparative functional genomic studies rely on pairwise comparisons, where traits of each gene are compared to traits of other genes across species. This leads to many more comparisons than unique observations, making each comparison dependent on others. (b) Comparative phylogenetic methods, including phylogenetic independent contrasts[2], make a smaller number of independent comparisons, where each contrast measures independent changes along different branches. Phylogenetic approaches are rarely used for functional genomic studies.

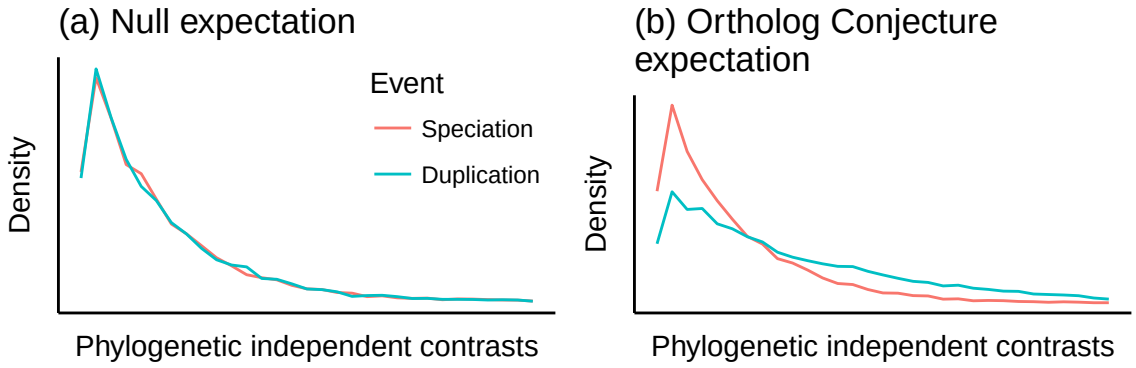

Figure 2: (a) Under the null hypothesis that there is no difference in the rate of evolution after duplication or speciation events, the distribution of phylogenetic independent contrasts would be drawn from the same distribution across both types of nodes. (b) Under the ortholog conjecture, contrasts across duplication nodes would tend to be larger than contrasts across speciation nodes.

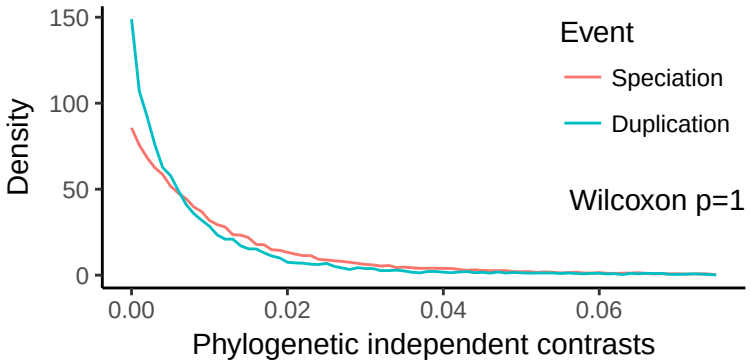

Figure 3: Phylogenetic reanalysis of the Tau values calculated by KMRR[12]. Density plot of the magnitude of phylogenetic independent contrasts for Tau following duplication and speciation events. As the contrasts for duplication events are not larger than those of speciation events, the results are not consistent with the predictions of the ortholog conjecture.

The KMRR study considered several publicly available datasets of gene expression across tissues and species. Their expression summary statistic is Tau[21], an indicator of tissue specificity of gene expression. It can range from a value of 0, which indicates no specificity (*i.e.*, uniform expression across tissues), to a value of 1, which indicates high specificity (*i.e.*, expression in only one tissue). Tau is convenient in that it is a single number of defined range for each gene, though of course since the original expression is multidimensional this means much information is discarded. This includes information about which tissue expression is specific to. For example, if one gene has expression specific to the brain and another expression specific to the kidney, both would have a Tau of 1.

The KMRR analyses are based on pairwise comparisons (Figure 1a) between Tau within each gene family. They found the correlation coefficient of Tau for orthologs to be significantly greater than the correlation coefficient of Tau for paralogs, *i.e.* that orthologs tend to be more similar to each other than paralogs. From this they concluded that their analyses support the ortholog conjecture. They also concluded that this pattern provides support for a particular evolutionary process, that "tissue-specificity evolves very slowly in the absence of duplication, while immediately after duplication the new gene copy differs"[12].

Phylogenetic reanalyses

We reanalyzed the KMRR study using phylogenetic comparative methods. We focused on one of the datasets included in their analyses, that of Brawand *et al.* 2011[23]. This dataset is the best sampled in their analyses. It has gene expression data for six organs across ten species (nine mammals and one bird), eight of which were analyzed by KMRR and further considered here.

For each internal node in each gene tree, we calculated the phylogenetic independent contrast (PIC) of Tau. This is the difference in values of Tau for descendant nodes scaled by the expected variance, which is largely determined by the lengths of the two branches that connect the node to its two descendants (Figure 1b). These contrasts were then annotated by whether each is made across a speciation or duplication event. The original description of independent contrasts[2] focused on assessing covariance between changes in two traits. Our use of contrasts is a bit different– we look for differences in evolutionary changes of one trait (differential expression) between two categories of nodes (speciation and duplication).

We mapped the Tau values calculated by KMRR[12] for the Brawand *et al.* 2011[23] datset onto 21124 gene trees parsed from ENSMBL Compara[24]. These are the same pre-computed trees that the orthology/ paralogy annotations KMRR used are based on. 8859 gene trees passed taxon sampling criteria (4 genes) after removing tips without Tau values and had at least one speciation event. Of these, 4763 were successfully time calibrated. These calibrated trees were used to calculate phylogenetic independent contrasts for 12856 duplication nodes and 43587 speciation nodes. One of these trees is shown in Supplementary Figure 7 to demonstrate the

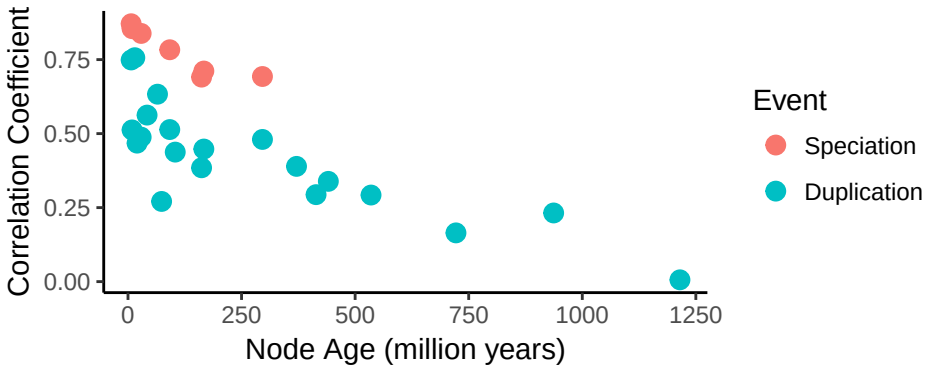

Figure 4: The pairwise analyses presented in Figure 2a of KMRR[12] can be reproduced with the subset of data we consider here. Each circle indicates the Tau Pearson correlation coefficient for a set of pairwise comparisons annotated with a specific node name of a given age and event type, *i.e.*, whether the divergence at that nodes was due to speciation (giving rise to orthologs) or duplication (giving rise to paralogs). Following KMRR Figure 2a, this figure shows only the comparisons that include a human sequence.

analysis.

It is essential to have a null hypothesis that makes a distinct prediction from the prediction of the hypothesis under consideration. A suitable null hypothesis in this case is that there is no difference in the evolution of expression following speciation or duplication events[25]. Under this hypothesis, we would predict that contrasts across speciation nodes and duplication nodes are drawn from the same distribution (Figure 2a). Under the alternative hypothesis specified by the ortholog conjecture, that there is a higher rate of change following duplication events than speciation events, we would expect to see the distribution of duplication contrasts shifted to higher values relative to the speciation contrasts (Figure 2b).

We did not find increased evolutionary change in expression following duplication events (Figure 3). The Wilcoxon rank test does not reject the null hypothesis that the rate of evolution following duplications is the same as or less than the rate following speciation (p value = 1). Our phylogenetic comparative analysis, unlike the previously published pairwise comparative analysis[12], therefore finds no evidence for the ortholog conjecture in this system.

We next examined the possibility that ascertainment biases were differentially impacting the inference of expression evolution following duplication and speciation events. Such a bias might obscure support for the ortholog conjecture. We focused on two possible sources of bias, node depth and branch length. We found no evidence that either affected our results (Supplementary Information, Supplementary Figure 8). We also examined the sensitivity of the results to the calibration times applied to speciation events on the gene trees. This is important because it is expected that genes from separate species have a common ancestor older than the time at which the species diverged from each other[26,27]. There is also uncertainty associated with the timing of these speciation events. We added random noise to the calibration times in replicate analyses, and all still failed to reject the null hypothesis (Supplementary Information).

Understanding the incongruence between pairwise and phylogenetic methods

In order to better understand why our phylogenetic analysis supports a different conclusion (*i.e.*, no support for the ortholog conjecture) than the published analysis of KMRR[12] (*i.e.*, strong support for the ortholog conjecture), we first checked to make sure we could reproduce their result based on pairwise analyses. This is important since we are only looking at at a subset of the data they considered, the Brawand *et al.* 2011[23] dataset for gene trees that could be successfully time calibrated. In their Figure 1, they present a higher Tau correlation coefficient between ortholog pairs than paralog ortholog pairs. We find the same here, with correlation coefficients of 0.77 for orthologs and 0.4 for paralogs.

Why is it that pairwise methods and phylogenetic methods lead to opposite conclusions? One reason

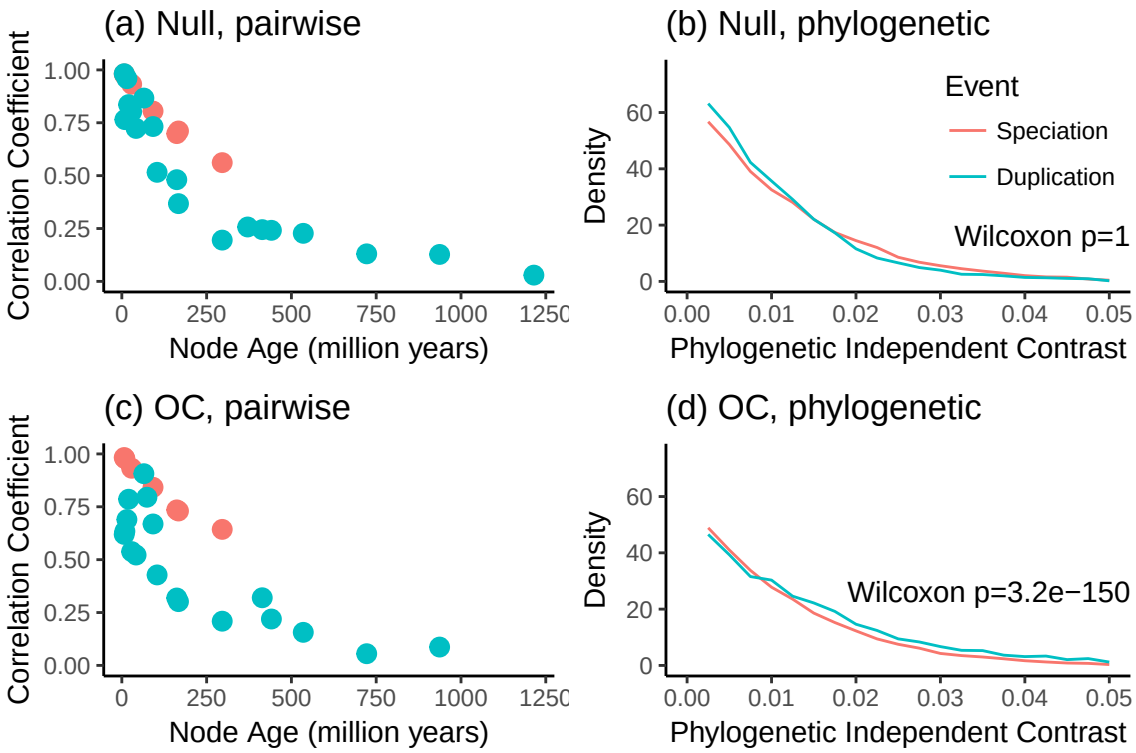

Figure 5: Analyses of simulated data reveal that pairwise methods produce similar results under both the null model and the ortholog conjecture (OC), while the phylogenetic method provides different results under each model. (a) Pairwise analysis of dataset simulated under the null model that there is no difference in the rate of expression evolution following duplication and speciation. (b) Phylogenetic analysis of data simulated under the null model. (c) Pairwise analysis of dataset simulated under the ortholog conjecture, in which there is a higher rate of expression evolution following duplication than speciation. This plot is very similar to the corresponding plot for the null model. (d) Phylogenetic analysis of data simulated under the ortholog conjecture. This plot is distinct from that for the null model in that phylogenetic independent contrasts for duplication events are significantly larger than after speciation events.

is that multiple pairwise comparisons repeatedly sample the same evolutionary changes and in so doing violate statistical assumptions of independence, whereas phylogenetic comparative methods make multiple independent comparisons across non-overlapping branches of the tree. The other reason for the different conclusions is that pairwise comparisons and phylogenetic comparative methods describe different things. Pairwise comparisons describe contemporary patterns, while phylogenetic methods infer historical processes[10]. There need not be a different process of evolution following speciation and duplication for paralogs to be more different than orthologs. Any difference could be due to the structure of the gene phylogenies alone. If paralogs tend to be more distantly related to each other than orthologs, then there would be more time for differences to accumulate even if the rate of change is the same between the two. This is, in fact, the case for these data. While the mean distance (*i.e.*, total branch length) between orthologs is 305.3 million years, the mean distance between paralogs is 1539.7 million years. This is because the oldest speciation event is by definition the most recent common ancestor of the species included in the study, but many gene families underwent duplication before this time.

To test the hypothesis that ancient duplications that precede the oldest speciation event (Supplementary Figure 8a) impact the lower correlation of Tau between paralogs than between orthologs, we removed them. When we consider only the duplication events the same age or younger than the oldest speciation event (Supplementary Figure 8b), the paralog correlation coefficient increases from 0.4 to 0.67. This is much closer to the ortholog correlation of 0.77.

The KMRR study did investigate the impact of node age on correlation, but in a different way. In their Figure 2, they grouped orthologs and paralogs according to the ENSEMBL node name of their most recent common ancestor, and plotted the correlation of Tau for each of these groups by the node age. They found that across the investigated range of node ages, ortholog pairs have higher Tau correlation than paralogs. We confirmed that we can replicate this result (Figure 4). There are several difficulties with interpreting this plot, though. First, it does not just reflect the evolutionary processes that generated the data, it is also impacted by the phylogenies along which these processes acted. The expected covariance of traits that evolve under neutral processes is in fact defined by the phylogeny[3]. Second, the correlation for each group is based on multiple non-independent pairwise comparisons.

To better understand this plot (Figure 4), we performed a couple simulations of Tau on the calibrated gene trees and then regenerated the figures. We did not modify the gene tree topologies or their inferred histories of duplication and loss. First, we simulated the evolution of Tau under the null model that it evolves at the same rate following duplication and speciation events. Under the null model, the plot of correlation coefficient to node age is very similar as for the observed data (Figure 5a). As in the original study, there is higher correlation coefficient across orthologs (0.74) than paralogs (0.28) when not considering node age. Phylogenetic analysis of the data simulated under the null hypothesis (Figure 5b) do not reject the null hypothesis (Wilcoxon $p = 1$), as expected.

We next simulated the evolution of Tau under the ortholog conjecture, where the rate of evolution of Tau following duplication was 5 fold the rate following speciation. The pairwise results of this heterogeneous model are nearly indistinguishable from the null model (Figure 5c), and also have a higher correlation coefficient for orthologs (0.78) than paralogs (0.2). The phylogenetic analysis of the ortholog conjecture simulation (Figure 5d) does reject the null hypothesis (Wilcoxon $p = 3.2 \times 10^{-150}$).

These simulations have several implications. The pairwise comparisons used by KMRR cannot distinguish between the null hypothesis and ortholog conjecture even when the evolution of expression is 5 fold higher following duplication than speciation. The pairwise results are strikingly similar under both hypotheses. These simulations also serve as a validation of phylogenetic methods as applied to this problem. Our phylogenetic analysis of independent contrasts does not reject the null hypothesis when data are simulated under the null model, and does reject the null hypothesis when the data are simulated under the ortholog conjecture. In contrast to pairwise methods, the phylogenetic analyses make testable predictions based on explicit hypotheses about evolutionary process.

Their Figure 3 differs from their other analyses in that it presents independent changes in expression between triplets of genes. Each triplet has two paralogs that arose from a duplication event and one unduplicated homolog. They find that there is a tendency for the paralog with the lowest expression to have the highest Tau. From this they conclude that following duplication there is a common trajectory of evolutionary change, where one paralog evolves to have lower expression and also become more tissue specific. There is, however, a negative relationship between Tau and maximum expression across all genes (Supplementary Figure 9). A simpler explanation for this plot is that it reflects the global tendency for Tau to decrease with maximum expression. The null expectation for any two genes sampled at random is for one to have higher maximum expression and lower Tau, and the other to have lower maximum expression and higher Tau. This global pattern could be due to both biological factors and the technical details of how Tau is calculated.

Implications for the ortholog conjecture

There has been considerable recent interest in, and controversy about, the ortholog conjecture[13,14,25,28,29]. While some studies have presented support for the ortholog conjecture, our results are consistent with multiple studies that have not[13,25,30].

Our results suggest, at a minimum, that the ortholog conjecture is not a dominant pattern that is central to explaining the evolution of phenotypic diversity in gene families. The ortholog conjecture does not have to be an all or nothing question, though. It may be the case that the rates of phenotypic evolution following duplication may be greater than that following duplication in some organisms, gene families, and evolutionary processes[28]. We just don't find evidence for it when summarizing gene expression across tissues with Tau in

these organisms. This calls into question the general predictive power of the ortholog conjecture with respect to gene expression, and until these processes are better understood it will be necessary to test for it in each situation. These tests should be articulated in terms of clear alternative hypotheses[25] that make distinct phylogenetic comparative predictions.

Lack of support for the ortholog conjecture has important biological implications. It indicates that the mechanism of gene divergence (speciation versus duplication) may not have as strong an impact on phenotypic divergence as sometimes proposed. It also has technical implications. Having information on whether two genes are orthologs or paralogs provides little added information about expression beyond knowing how distantly related the two genes are. Rather than focus on whether genes are orthologs or paralogs when attempting to predict function, it may be more effective to simply focus on how closely related or distantly related they are. Closely related paralogs, for example, may tend to have more similar phenotypes than more distantly related orthologs[25].

This is also an example of the limitations of the concepts of orthology and paralogy[31]. These terms can have straightforward meaning in small gene trees with simple duplication/speciation histories, but the utility of the terms breaks down on larger more complex gene trees. Orthology and paralogy are annotations on the tips of the phylogeny that are derived from the structure of the tree and history of duplication and speciation at internal tree nodes. In this sense, orthology and paralogy are statements about the internals of the tree that are distilled into statements at the tips of the tree. Much is lost in the process, though. For most questions it is much more direct to focus on the structure of the tree and the inferred processes within the tree, such as which internal nodes are duplication or speciation events and how much change occurs along the branches.

Levin *et al.* reanalysis

Original pairwise analyses of developmental gene expression

Levin *et al.*[15] analyzed gene expression through the course of embryonic development for ten animal species, each from a different clade that has been designated as having the rank of phylum. They arrived at two major conclusions. First, animal development is characterized by a well-defined mid-developmental transition that marks the transition from an early phase of gene expression to a late stage of gene expression. Second, this transition helps explain the evolution of features observed among distantly related animals. Specifically, they concluded that animals from different phyla exhibit an "inverse hourglass" model for the evolution of gene expression, where there is more evolutionary variance in gene expression at a mid phase of development than there is at early and late phases. Closely related animals have previously been described as having an hourglass model of gene expression, where evolutionary variance in expression is greater early and late in development than at the midpoint of development[16,32]. Levin *et al.* conclude that this contrast between distantly and closely related animals provides biological justification for the concept of phyla and may provide a definition of phyla.

Levin *et al.*[15] arrived at this conclusion by making multiple pairwise comparisons of ortholog expression data sampled throughout the course of embryonic development. For each species pair, they identified the orthologs shared by these species. This list of shared genes was different from species pair to species pair. They characterized each of these orthologs in each species as having expression that peaks in early, mid, or late temporal phase of development. They then calculated a similarity score for each temporal phase for each species pair based on the fraction of genes that exhibited the same patterns in each species. The distributions of similarity scores are plotted in their Figure 4d, and their Kolmogorov–Smirnov (KS) tests indicated that the early distribution and late distribution were each significantly different from mid distribution ($P < 10^{-6}$ and $P < 10^{-12}$, respectively). This is the support they presented for the inverse hourglass model.

Reexamination of pairwise comparisons

We examined the matrix of pairwise comparisons used as the base for the KS tests and Figure 4d in *Levin et al.*[15], and thus as evidence to support the "inverse hourglass" model. We found several problems resulting from the use of multiple pairwise comparisons. The first problems are specific to this particular implementation of

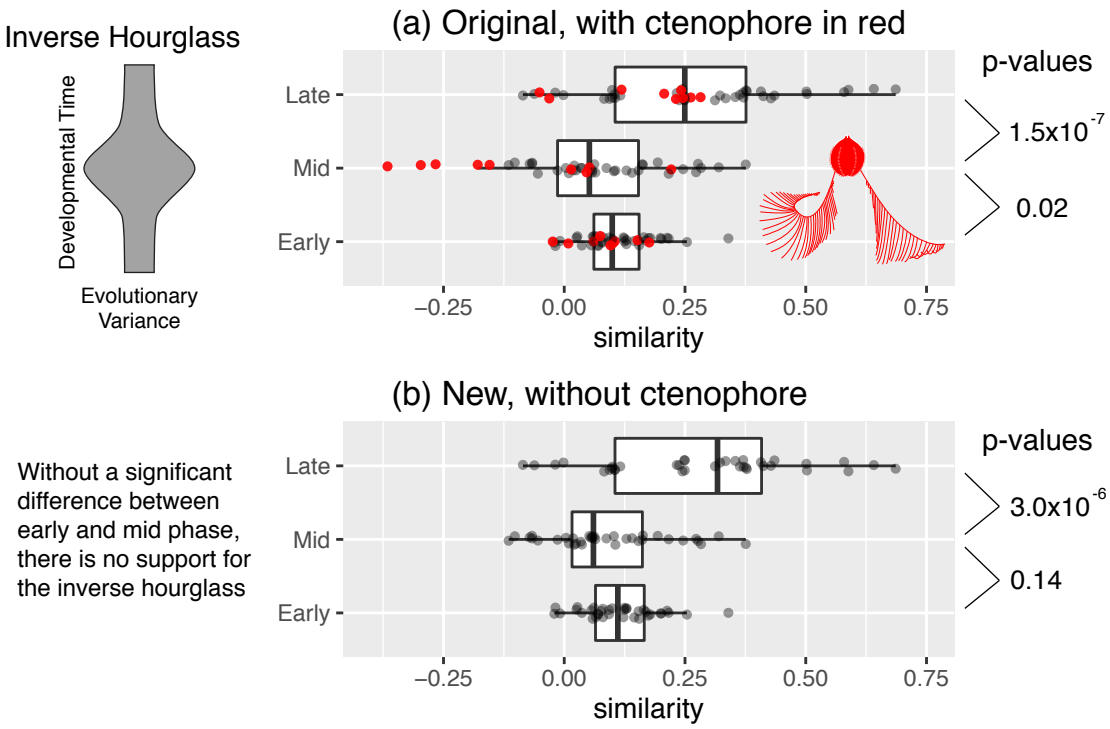

Figure 6: Distributions of pairwise similarity scores for each phase of development. Pairwise scores for the ctenophore are red. Wilcoxon test p-values for the significance of the differences between early-mid distributions and late-mid distributions are on the right. Model of variance, which is inversely related to similarity, is on the left. (a) The distributions as published by Levin *et al.*[15]. Low similarity (*i.e.*, high variance) in the mid phase of development was interpreted as support for an inverse hourglass model for the evolution of gene expression. The five least-similar mid phase scores were all from the ctenophore. Published KS p-values, based on duplicated data, are in parentheses. The inset ctenophore image is by S. Haddock from phylopic.org. (b) The distributions after the exclusion of the ctenophore. The early and mid phase distributions are not statistically distinct.

pairwise comparisons. We found that every data point was included twice because both reciprocal pairwise comparisons (which have the same values) were retained. For example, there is both a nematode to arthropod comparison and an arthropod to nematode comparison. As a consequence, there are 90 entries for the 45 pairwise comparisons, and by doubling the data the significance of the result appears stronger than it actually is. After removing the duplicate values, the p values are far less significant, 0.002 for the early-mid comparison and on the order of 10^{-6} for early-late. In addition, the test they used (KS test) is not appropriate for the hypothesis they seek to evaluate. The KS test does not just evaluate whether one distribution is greater than the other, it also tests whether the shape of the distributions are the same. In addition, the samples in this dataset are matched (*i.e.*, for each pairwise comparison there is a early, mid, and late expression value), which the KS test does not take into account. The Wilcoxon test is instead appropriate in this case. When applied to the de-duplicated data, the significance of this test is 0.02 for the early-mid comparison and on the order of 10^{-7} for early-late.

Once we addressed the issues above with the implementation of pairwise comparisons, we were able to explore more general issues that can be a problem when making multiple pairwise comparisons between species. We found that all five of the lowest values in the mid phase distribution (Figure 6a) are for pairwise comparisons that include the ctenophore (comb jelly). When the nine pairwise comparisons that include the ctenophore are removed, there is no significant difference between the early phase and mid phase distributions ($p = 0.14$ for the early-mid comparison and $p < 10^{-5}$ for the late-mid comparison) and no support for the inverse hourglass (Figure 6b). This highlights a well understood property of pairwise comparisons across species[2,11]: evolutionary changes along a given branch, like those along the ctenophore branch, impact each of the multiple pairwise comparisons that includes that branch. The pairwise comparisons are therefore not independent - different pairwise comparisons are impacted by changes along some of the same branches (Figure 1a). This can give the impression of a general pattern across the tree that is instead specific to changes along one part of the tree. The number of comparisons impacted by each change depends on the structure of the phylogenetic tree, *i.e.* how the species are related to each other. Phylogenetic comparative methods were developed specifically to address this problem[2].

While we demonstrate problems with pairwise comparisons that impacts the Levin *et al.* analysis, we did not perform a phylogenetic reanalysis of this study, as we did for the KMRR study. This is because the similarity metric computed in the pairwise comparisons of Levin *et al.* is not suitable for phylogenetic analysis. One issue with the similarity metric is that it is based on different genes for different species pairs, and therefore is not a trait whose evolution can be modeled across the phylogeny. A full phylogenetic reanalysis would be possible using upstream analysis products to re-derive new expression summary statistics.

Phylogenetic comparative methods in functional genomics

Our results highlight the importance of explicitly incorporating information about phylogenetic relationships when comparing functional genomic traits across species. Some of the most widely used phylogenetic comparative methods[2,3] are already directly applicable to comparative functional genomics studies. There are also interesting new challenges at this interdisciplinary interface that will need to be addressed to fully realize the potential of phylogenetic comparative functional genomics studies. One such challenge is that most phylogenetic comparative analyses of covariance between traits have been developed to address problems with many more species (e.g., dozens or more) relative to the number of traits being examined. In comparative functional genomics analyses, there are often far fewer species because adding taxa is still expensive, but tens of thousands of traits. This creates statistical challenges as the resulting covariance matrices are singular and, if not treated appropriately, imply many false correlations that are artifacts of project design. We outlined these challenges and potential solutions in the context of gene expression[33]. In the same manuscript we also considered another issue of relevance here – the read counts generated by RNA-seq expression studies cannot be directly compared across species. This is because there are various species-specific technical factors that can be mistaken for differences in expression across species. These can be canceled out within species before making comparisons across species[33].

Recent advances in phylogenetic comparative methods are particularly well suited to addressing questions about the evolution of functional genomic traits. Most early phylogenetic comparative methods attempted to

account for evolutionary signal to correct statistical tests for correlations between traits, while more recent methods tend to focus on testing hypotheses of evolutionary processes[34]. The application of this newer focus to functional genomics provides an exciting opportunity to address long standing questions of broad interest, including the order of changes in functional genomics traits and shifts in rates of evolution of one functional genomics trait following changes in another trait.

We are not the first to apply phylogenetic comparative methods to functional genomic data. While the vast majority of comparative functional genomics studies have used standard pairwise similarity methods, a small number of comparative functional genomics studies have employed phylogenetic comparative approaches[35–37]. For instance, a phylogenetic ANOVA[38] of the evolution of gene expression improves statistical power and drastically reduces the rate of false positives relative to pairwise approaches.

Addressing the statistical dependence of pairwise comparisons is not the only advantage of using phylogenetic comparative methods for functional genomics analyses. Another problem with the pairwise comparisons is that, except at the tips, they summarize changes along many branches in the phylogeny. Two paralogs that diverged from a duplication event deep in the tree may have many subsequent duplication and speciation events, and changes along all these branches will impact the final pairwise comparison. Phylogenetic methods have the advantage of isolating the changes under consideration (Figure 1b). The phylogenetic methods avoid diluting the change that occurs along the branches that follow the node in question with changes along all subsequent branches. There may still be missing speciation events, due to extinction and incomplete taxon sampling, and missing duplication events, due to gene loss, but these omissions affect both methods.

Conclusions

The fact that the first two comparative functional genomics studies we reanalyzed show serious problems with pairwise comparisons indicates that there likely to be similar problems in other studies that use these methods, and that future comparative studies will be compromised if they continue to use pairwise methods. Studies of evolutionary functional genomics should not be focused on the tips of the tree using pairwise comparisons. They should explicitly delve into the tree with phylogenetic comparative methods. Here we show a concrete case where the predictions of a null hypotheses and an alternatives hypothesis do not differ when considering pairwise comparisons, but make distinct predictions with phylogenetic comparative methods that can be used to test them.

These analyses illustrate how important it is to not conflate evolutionary patterns with the processes that generated them. Finding a pattern wherein paralogs tend to be more different than orthologs is not evidence that there are different processes by which orthologs and paralogs evolve. This is also the expected pattern when they evolve under the same process but paralogs tend to be more distantly related to each other than orthologs are. The fact that multiple pairwise comparisons of developmental gene expression across diverse species share a particular pattern is not evidence of a general process that explains the differences between all species in the analysis. It is also the expected pattern when a single species has unique differences, and the evolutionary changes responsible for these differences are sampled multiple times in pairwise comparisons that span the same phylogenetic branches along which these differences arose. To use patterns across living species to test hypotheses about evolutionary processes it is also necessary to incorporate information about evolutionary relationships, *i.e.* phylogenies. There have been decades of work on building comparative phylogenetic methods that do exactly that, and they are just as relevant to comparing functional genomics traits across species as they are to comparing morphology or any of the other traits they are already routinely applied to.

Methods

All files needed to re-execute the analyses presented in this document are available at https://github.com/caseywdunn/comparative_expression_2017. The most recent commit at the time of the analysis presented here was 6b3639bc4e097402ea4dce51c33625800b8e8ca0. See the `readme.md` file in this repository for more information on the contents of the source file and how to re-execute them.

KMRR reanalysis

The KMRR study[12] followed excellent practices in reproducibility. They posted all data and code needed to re-execute their analyses at figshare: https://figshare.com/articles/Tissue-specificity_of_gene_expression_diverges_slowly_between_orthologs_and_rapidly_between_paralogs/3493010/2 . We slightly altered their `Rscript.R` to simplify file paths and specify one missing variable. This modified script and their data files are available in the github repository for this paper, as are the intermediate files that were generated by their analysis script that we used in our own analyses. We also obtained the `Compara.75.protein.nh.emf` gene trees[24] from ftp://ftp.ensembl.org/pub/release-75/emf/ensembl-compara/homologies/ and place them in the same directory as this file. These gene trees include branch lengths and annotate each internal node as being a duplication or speciation event.

We considered only the data from Brawand *et al.* 2011[23] for the eight taxa included in KMRR. We left in sex chromosome genes and testes expression data, which KMRR removed in some of their sensitivity analyses. This corresponded to the analyses that provided the strongest support for the ortholog conjecture and therefore the most conservative reconsideration of it.

After parsing the trees from the Compara file with treeio, which was recently split from ggtree[39], we added Tau estimates generated by the KMRR `Rscript.R` to the tree data objects. We then pruned away tips without expression data, retaining only the trees with 4 or more tips. We also only retained trees with one or more speciation events, as speciation events are required for calibration steps. This removes trees that have multiple genes from only one species after pruning away tips without expression data.

The gene trees were then time calibrated. The goal is not necessarily to have precise dates for each node, but to scale branch lengths so that they are equivalent across the trees. This in turn scales the phylogenetic independent contrasts (which take branch length into account) so they can be compared appropriately. We found a problem with the clade names on the ENSEMBL Compara trees, which others have also identified[40], that first had to be fixed. Hominini is the name for the clade that includes humans and chimps, while Homininae is the clade that includes humans, chimps, and gorillas. Both clades are labeled as Homininae in the Compara trees, though, which would interfere with calibration of these nodes. To fix the problem, we identified all clades labeled Homininae that have no gorilla sequence and renamed them Hominini. We then calibrated the trees by fixing the speciation nodes to the dates specified in the KMRR code, with the exception of Hominini and Homininae. These we set to 7 million years and 9 million years, drawing on the same TimeTree source[41] that KMRR used. We used the `chronos()` function from the R package ape[42] for this calibration, with the `correlated` model. See Supplementary Information for additional sensitivity analyses to time calibration. Some trees could not be calibrated with these hard node constraints, and were discarded.

For each node in the remaining calibrated trees, we calculated the phylogenetic independent contrast for Tau across its daughter branches with the `pic()` function in ape[42]. We then collected the contrasts from all trees into a single table, along with other annotations including whether the node is a speciation or duplication event. This table, `nodes_contrast`, was then analyzed as described in the main text for the presented plots and tests.

Levin et al. reanalysis

Levin *et al.* helpfully provided data and clarification on methods. We obtained the matrix of pairwise scores that underlies their Figure 4d and confirmed we could reproduce their published results. We then removed duplicate rows, applied the Wilcoxon test in place of the Kolmogorov-Smirnov test, and identified ctenophores as overrepresented among the low outliers in the mid-developmental transition column. An annotated explanation of these analyses is included in the git repository at https://github.com/caseywdunn/comparative_expression_2017/blob/master/levin_etal/reanalyses.md .

Acknowledgments

Thanks to Steve Haddock, Alex Damian Serrano, August Guang, Bruno Vellutini, and Zack Lewis for feedback on the manuscript. CWD's contribution was supported by the National Science Foundation (DEB-1256695 and the Waterman Award). AH was supported by the European Research Council Community's Framework Program Horizon 2020 (2014–2020) ERC grant agreement 648861.

Supplementary Information

KMRR Analyses

Summary statistics

Table 1: Number of speciation nodes N with contrasts by clade name.

Clade	N
Euarchontoglires	4096
Mammalia	6038
Homininae	8297
Amniota	5825
Catarrhini	7413
Theria	6158
Hominini	5760

Table 2: Number of tips for each species in trees that were used to calculate the independent contrasts.

Species	N
Gallus gallus	7056
Gorilla gorilla	8266
Homo sapiens	8549
Macaca mulatta	9162
Monodelphis domestica	8254
Mus musculus	8733
Ornithorhynchus anatinus	9637
Pan troglodytes	7983

Investigation of potential ascertainment bias

While the age of speciation nodes is constrained, duplication nodes can be much older and therefore have a wider range of ages (Supplementary Figure 8a). This is because many gene duplication events are older than the most recent common ancestor of the species in the study. There are also technical factors that can lead to an excess of duplication events deeper in the tree. Gene tree estimation errors errors, for example, tend to lead to the overestimation of deep duplications[43]. If independent contrast values also tended to to be lower at greater node depth, it could interact with the preponderance of duplications at greater depth to create a pattern of lower contrasts associated with duplication events. To test for such an effect, we remove duplication nodes that are older than the oldest speciation node. Our results are unchanged and this reduced dataset does not reject the null hypothesis that the rate of evolution following duplications is the same as or less than the rate following speciation (Supplementary Figure 8b).

The independent contrast across a node is the amount of change observed between the daughter notes, scaled by the expected variance[2]. The expected variance is principally determined by the lengths of the branches leading from the node to these daughters. The shorter the total length of the two branches leading to daughter nodes, the larger the contrast for a given observed difference. This is because the same difference across a shorter total branch length indicates a greater rate of evolutionary change. The expected variance of contrasts for speciation nodes is constrained by the branch lengths on the species tree, but the expected variance of contrasts for duplications has a much wider range (Figure 8c). This could lead to biases if the lengths of

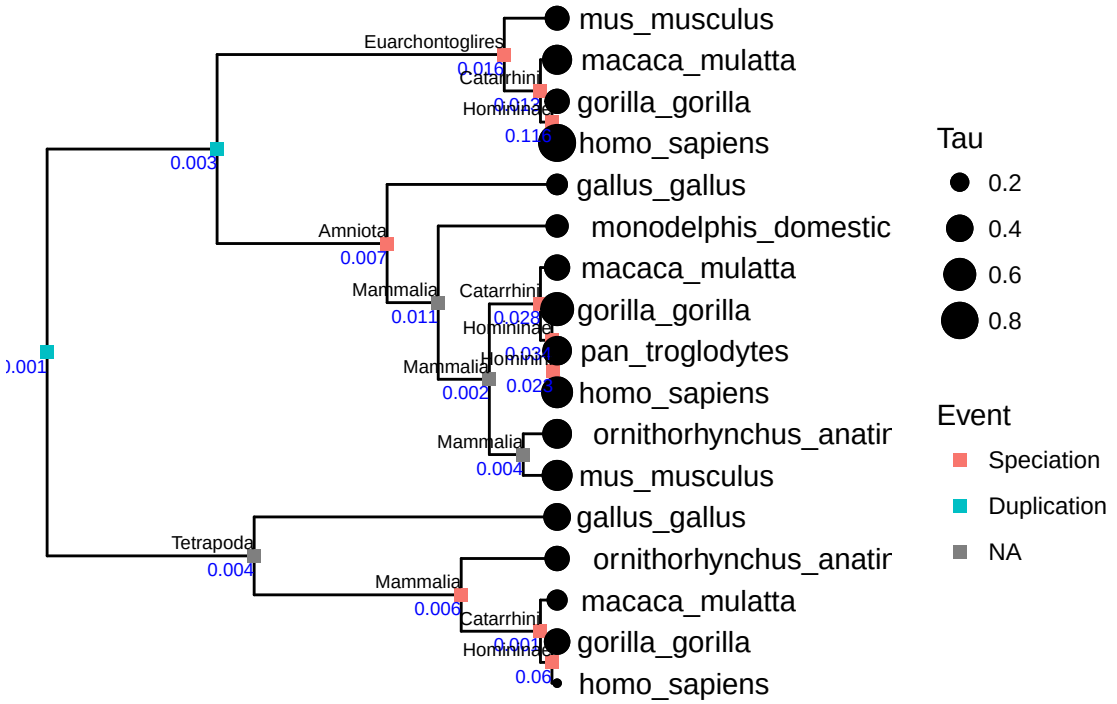

Figure 7: One of the gene trees from the phylogenetic analysis. The value of Tau (expression specificity) is indicated by the sizes of the circles at the tips of the tree. Whether an internal node is a speciation or duplication is indicated by color. Speciation nodes are labeled by clade name. Branch lengths are scaled to time. The blue number is the independent contrast for each node.

branches that descend from duplication nodes tend to be overestimated. We therefore examined only the contrasts that fell within in the range of expected variance seen for speciation contrasts, excluding duplication contrasts that fall outside of this range. This reanalysis does not reject the null hypothesis either (Figure 8d), indicating that branch length bias is not responsible for the result.

Investigation of sensitivity to calibration times

We examined the sensitivity of our results to the specification of calibration dates for the speciation nodes. In 10 reanalyses, we drew a new date for each calibration from a normal distribution with the mean of the original date and a standard deviation 0.2 times the original date. If any daughter nodes became older than their parent, we repeated the sampling until the dates were congruent with the topology. The minimum Wilcoxon p in these reanalyses was 1, *i.e.* none of them reject the null hypothesis that the rate of evolution of Tau is greater following duplication events than speciation events. This is consistent with the analysis that uses the calibration dates as specified, indicating that our results are robust to the selection of calibration times for speciation nodes.

Relationship between Tau and maximum expression

There is a negative relationship between miximum expression and Tau for each gene (Supplementary Figure 9).

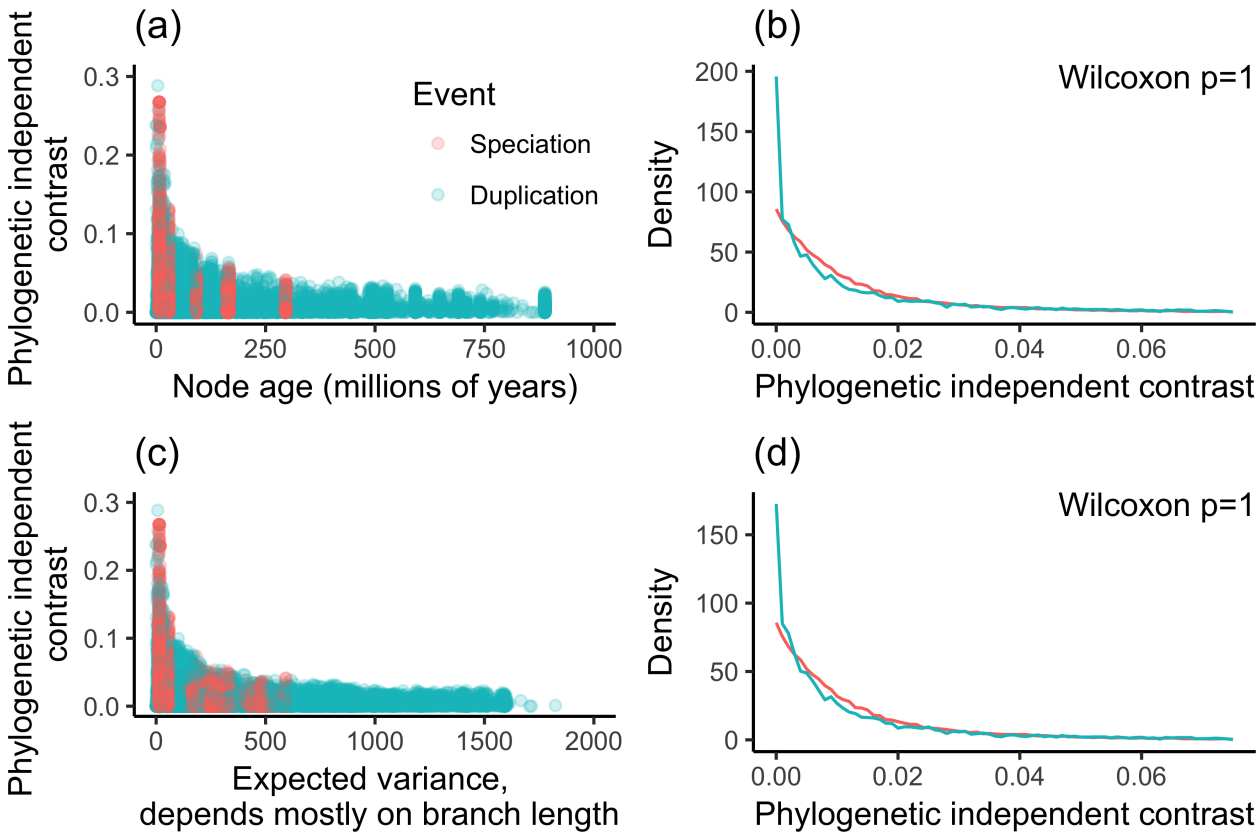

Figure 8: Investigation of possible ascertainment biases. (a) Magnitude of independent contrasts plotted against node age. Speciation nodes are calibrated to particular times, whereas duplication nodes have a wider range. (b) Density plot of contrasts for only the nodes that have an age less than or equal to the maximum age of speciation nodes. (c) Magnitude of independent contrasts plotted against expected variance, which is largely determined by branch lengths. Contrasts for speciation nodes have a narrower range of expected variance than do contrasts for duplication nodes. (d) Density plot of contrasts for only the nodes that have expected variance within the range of contrasts across speciation nodes.

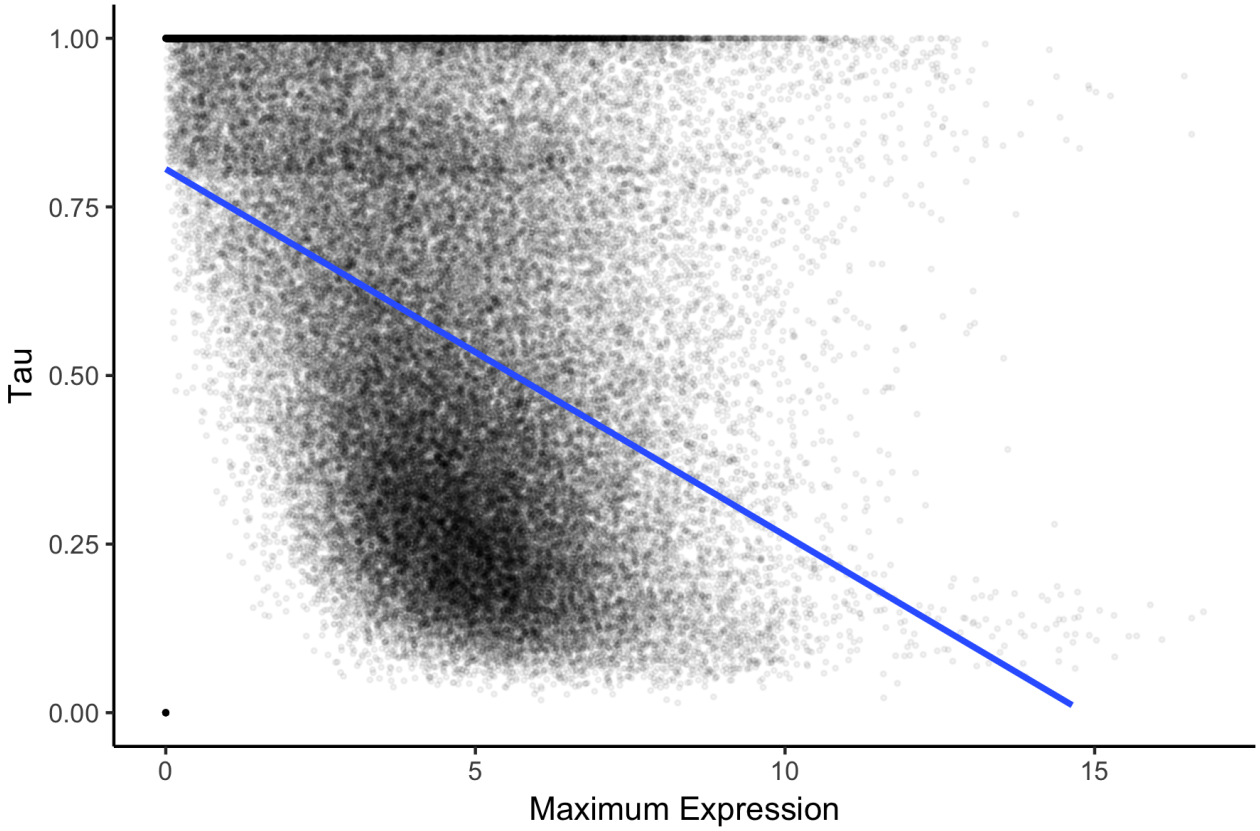

Figure 9: Genes with higher maximum observed expression tend to have lower Tau. The blue line is the linear model for the relationship between the two.

Software versions

This manuscript was computed on Sat Feb 11 01:16:55 2017 with the following R package versions.

```
R version 3.3.2 (2016-10-31)
Platform: x86_64-apple-darwin13.4.0 (64-bit)
Running under: macOS Sierra 10.12.3

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] parallel  stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
 [1] phytools_0.5-64  maps_3.1.1       geiger_2.0.6     gridExtra_2.2.1
 [5] hutan_0.0.0.9000 digest_0.6.11    ape_4.0          ggtree_1.7.7
 [9] stringr_1.1.0    magrittr_1.5     dplyr_0.5.0      purrr_0.2.2
[13] readr_1.0.0      tidyr_0.6.1      tibble_1.2       ggplot2_2.2.1
[17] tidyverse_1.0.0  treeio_0.99.9    devtools_1.12.0

loaded via a namespace (and not attached):
 [1] phangorn_2.1.1          deSolve_1.14
 [3] subplex_1.2-2           splines_3.3.2
 [5] lattice_0.20-34         colorspace_1.3-2
 [7] expm_0.999-0            htmltools_0.3.5
 [9] yaml_2.1.14             survival_2.40-1
[11] withr_1.0.2             DBI_0.5-1
[13] plyr_1.8.4              combinat_0.0-8
[15] munsell_0.4.3           gtable_0.2.0
[17] mvtnorm_1.0-5           codetools_0.2-15
[19] coda_0.19-1             memoise_1.0.0
[21] evaluate_0.10           labeling_0.3
[23] knitr_1.15.1            highr_0.6
[25] Rcpp_0.12.9.1           scales_0.4.1
[27] backports_1.0.4         plotrix_3.6-4
[29] clusterGeneration_1.3.4 scatterplot3d_0.3-38
[31] jsonlite_1.2            fastmatch_1.0-4
[33] mnormt_1.5-5            stringi_1.1.2
[35] msm_1.6.4               animation_2.4
[37] numDeriv_2016.8-1       grid_3.3.2
[39] rprojroot_1.1           quadprog_1.5-5
[41] tools_3.3.2             lazyeval_0.2.0
[43] MASS_7.3-45             Matrix_1.2-7.1
[45] assertthat_0.1          rmarkdown_1.3
[47] R6_2.2.0                igraph_1.0.1
[49] nlme_3.1-128
```

References

1. Wray, G. A. Genomics and the Evolution of Phenotypic Traits. *Annual Review of Ecology, Evolution, and Systematics* **44,** 51–72 (2013).

2. Felsenstein, J. Phylogenies and the Comparative Method. *American Naturalist* **125,** 1–15 (1985).

3. Grafen, A. The phylogenetic regression. *Philosophical Transactions of the Royal Society B: Biological Sciences* **326,** 119–157 (1989).

4. Pagel, M. Inferring the historical patterns of biological evolution. *Nature* **401,** 877–884 (1999).

5. FitzJohn, R. G. Diversitree: comparative phylogenetic analyses of diversification in R. *Methods in Ecology and Evolution* **3,** 1084–1092 (2012).

6. Uyeda, J. C. & Harmon, L. J. A Novel Bayesian Method for Inferring and Interpreting the Dynamics of Adaptive Landscapes from Phylogenetic Comparative Data. *Systematic Biology* **63,** 902–918 (2014).

7. Garamszegi, L. Z. *Modern Phylogenetic Comparative Methods and their Application in Evolutionary Biology.* (Springer Berlin Heidelberg, 2014). doi:10.1007/978-3-662-43550-2

8. Ricklefs, R. E. & Starck, J. M. Applications of Phylogenetically Independent Contrasts: A Mixed Progress Report. *Oikos* **77,** 167–172 (1996).

9. Chamberlain, S. A. *et al.* Does phylogeny matter? Assessing the impact of phylogenetic information in ecological meta-analysis. *Ecology Letters* **15,** 627–636 (2012).

10. O'Meara, B. C. Evolutionary Inferences from Phylogenies: A Review of Methods. *Annual Review of Ecology, Evolution, and Systematics* **43,** 267–285 (2012).

11. Ackerly, D. D. & Reich, P. B. Convergence and Correlations among Leaf Size and Function in Seed Plants: A Comparative Test Using Independent Contrasts. *American Journal of Botany* **86,** 1272–10 (1999).

12. Kryuchkova-Mostacci, N. & Robinson-Rechavi, M. Tissue-Specificity of Gene Expression Diverges Slowly between Orthologs, and Rapidly between Paralogs. *PLoS Computational Biology* **12,** e1005274–13 (2016).

13. Nehrt, N. L., Clark, W. T., Radivojac, P. & Hahn, M. W. Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals. *PLoS Computational Biology* **7,** e1002073 (2011).

14. Koonin, E. V. Orthologs, paralogs, and evolutionary genomics. *Annual Review of Genetics* **39,** 309–338 (2005).

15. Levin, M. *et al.* The mid-developmental transition and the evolution of animal body plans. *Nature* **531,** 637–641 (2016).

16. Kalinka, A. T. *et al.* Gene expression divergence recapitulates the developmental hourglass model. *Nature* **468,** 811–814 (2010).

17. Hejnol, A. & Dunn, C. W. Animal Evolution: Are Phyla Real? *Current Biology* **26,** R424–R426 (2016).

18. Chen, X. & Zhang, J. The Ortholog Conjecture Is Untestable by the Current Gene Ontology but Is Supported by RNA Sequencing Data. *PLoS Computational Biology* **8,** e1002784 (2012).

19. Thomas, P. D. *et al.* On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report. *PLoS Computational Biology* **8,** e1002386 (2012).

20. Forslund, K., Pekkari, I. & Sonnhammer, E. L. Domain architecture conservation in orthologs. *BMC Bioinformatics* **12,** 326 (2011).

21. Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21,** 650–659 (2005).

22. Kryuchkova-Mostacci, N. & Robinson-Rechavi, M. A benchmark of gene expression tissue-specificity

metrics. *Briefings in Bioinformatics* (2016). doi:10.1093/bib/bbw008

23. Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478,** 343–348 (2011).

24. Herrero, J. *et al.* Ensembl comparative genomics resources. *Database* **2016,** bav096–17 (2016).

25. Studer, R. A. & Robinson-Rechavi, M. How confident can we be that orthologs are similar, but paralogs differ? *Trends in Genetics* 1–7 (2009). doi:10.1016/j.tig.2009.03.004

26. Takahata, N. Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics* **122,** 957–966 (1989).

27. Degnan, J. H. & Rosenberg, N. A. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution* **24,** 332–340 (2009).

28. Gabaldon, T. & Koonin, E. V. Functional and evolutionary implications of gene orthology. *Nature Reviews Genetics* **14,** 360–366 (2013).

29. Altenhoff, A. M., Studer, R. A., Robinson-Rechavi, M. & Dessimoz, C. Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Computational Biology* **8,** e1002514 (2012).

30. Yanai, I., Graur, D. & Ophir, R. Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. *Omics : a journal of integrative biology* **8,** 15–24 (2004).

31. Dunn, C. W. & Munro, C. Comparative genomics and the diversity of life. *Zoologica Scripta* **45,** 5–13 (2016).

32. Domazet-Lošo, T. & Tautz, D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* **468,** 815–818 (2010).

33. Dunn, C. W., Luo, X. & Wu, Z. Phylogenetic analysis of gene expression. *Integrative and Comparative Biology* **53,** 847–856 (2013).

34. Pennell, M. W. & Harmon, L. J. An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. *Annals of the New York Academy of Sciences* **1289,** 90–105 (2013).

35. Oakley, T. H., Gu, Z., Abouheif, E., Patel, N. H. & Li, W.-H. Comparative methods for the analysis of gene-expression evolution: an example using yeast functional genomic data. *Molecular Biology and Evolution* **22,** 40–50 (2005).

36. Eng, K. H., Bravo, H. C. & Keleş, S. A phylogenetic mixture model for the evolution of gene expression. *Molecular Biology and Evolution* **26,** 2363–2372 (2009).

37. Chang, D. & Duda, T. F. Application of community phylogenetic approaches to understand gene expression: differential exploration of venom gene space in predatory marine gastropods. *BMC Evolutionary Biology* **14,** 123 (2014).

38. Rohlfs, R. V. & Nielsen, R. Phylogenetic ANOVA: The Expression Variance and Evolution Model for Quantitative Trait Evolution. *Systematic biology* **64,** 695–708 (2015).

39. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* (2016). doi:10.1111/2041-210X.12628

40. Daub, J., Moretti, S., Davydov, I. I., Excoffier, L. & Robinson-Rechavi, M. Detection of pathways affected by positive selection in primate lineages ancestral to humans. *bioRxiv* 044941 (2016). doi:10.1101/044941

41. Hedges, S. B., Dudley, J. & Kumar, S. TimeTree: a public knowledge-base of divergence times among

organisms. *Bioinformatics* **22,** 2971–2972 (2006).

42. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20,** 289–290 (2004).

43. Hahn, M. W. Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biology* **8,** R141 (2007).