Pairwise comparisons across species are problematic when analyzing functional genomic data

Casey W. Dunn[a,1,2], Felipe Zapata[b], Catriona Munro[a], Stefan Siebert[c], and Andreas Hejnol[d]

[a]Department of Ecology and Evolutionary Biology, Brown University, Providence, RI 02912; [b]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095; [c]Department of Molecular and Cellular Biology, University of California, Davis, CA 95616; and [d]Sars International Centre for Marine Molecular Biology, University of Bergen, Bergen 5006, Norway

Edited by David M. Hillis, The University of Texas at Austin, Austin, TX, and approved December 1, 2017 (received for review May 8, 2017)

There is considerable interest in comparing functional genomic data across species. One goal of such work is to provide an integrated understanding of genome and phenotype evolution. Most comparative functional genomic studies have relied on multiple pairwise comparisons 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, and find reason to question the original conclusions of both. One study interpreted pairwise comparisons of gene expression as support for the ortholog conjecture, the hypothesis that orthologs tend to have more similar attributes (expression in this case) than paralogs. The other study interpreted pairwise comparisons of embryonic gene expression across distantly related animals as evidence for a distinct evolutionary process that gave rise to phyla. In each study, distinct patterns of pairwise similarity among species were originally interpreted as evidence of particular evolutionary processes, but instead, we find that they reflect species relationships. These reanalyses concretely show the inadequacy of pairwise comparisons for analyzing functional genomic data across species. It will be critical to adopt phylogenetic comparative methods in future functional genomic work. Fortunately, phylogenetic comparative biology is also a rapidly advancing field with many methods that can be directly applied to functional genomic data.

phylogenetics | functional genomics | gene expression | ortholog conjecture | hourglass

The focus of genomic 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 critical to transferring functional insight across species and will grow in importance in coming years.

Over the last three decades, a rich set of phylogenetic comparative methods has been developed to address the challenges and opportunities of trait comparisons across species (2–10). A central challenge is the dependence of observations across species because of the evolutionary history of species—more closely related species share more traits, because they have a more recent 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 (11). Even so, most comparative functional genomic studies have abstained from phylogenetic approaches, and instead rely on multiple pairwise comparisons across species (Fig. 1*A*). This leaves comparative functional genomic 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 (Fig. 1*B*). The value of these methods relative to pairwise comparisons has been repeatedly shown in analyses of other types of character data (12–15). One reason that comparative functional genomic studies have not embraced phylogenetic approaches is that there has not yet been a concrete demonstration that pairwise and phylogenetic comparative methods can lead to different results when considering functional genomic data. Here, we examine this issue by reevaluating the pairwise comparisons in two recent studies that compared gene expression across species.

In the first study, Kryuchkova-Mostacci and Robinson-Rechavi (KMRR) (16) 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 paralogs (17,18). Using pairwise comparisons (Fig. 1*A*), they found lower expression correlation between paralogs than between orthologs and interpreted this as strong support for the ortholog conjecture.

The second comparative functional genomic study that we evaluate here is by Levin et al. (19). This study analyzed gene

Significance

Comparisons of genome function between species are providing important insight into the evolutionary origins of diversity. Here, we show that comparative functional genomics studies can come to the wrong conclusions if they do not take the relationships of species into account and instead rely on pairwise comparisons between species, as is common practice. We reexamined two previously published studies and found problems with pairwise comparisons that draw both their original conclusions into question. One study found support for the ortholog conjecture, and the other concluded that evolution of gene expression differed in pattern and process between animal phyla vs. within animal phyla. Our results show that, to answer evolutionary questions about genome function, it is critical to consider evolutionary relationships.

Author contributions: C.W.D., F.Z., C.M., S.S., and A.H. designed research; C.W.D. and F.Z. performed research; C.W.D. and F.Z. analyzed data; and C.W.D., F.Z., C.M., S.S., and A.H. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: casey.dunn@yale.edu.

[2]Present address: Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06511.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1707515115/-/DCSupplemental.

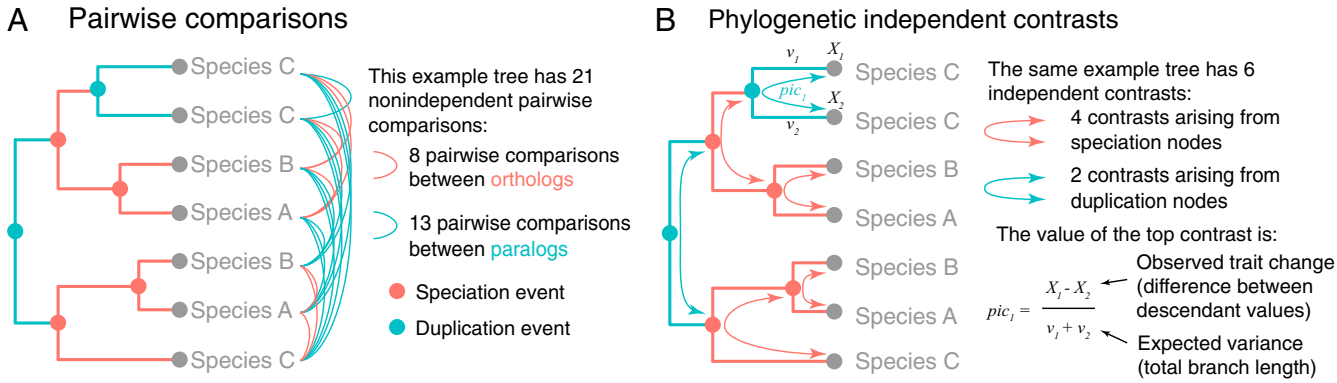

Fig. 1. Pairwise and phylogenetic comparative approaches illustrated on an example gene tree with multiple genes per species. The internal nodes of the tree are speciation and gene duplication events. (*A*) Many comparative functional genomic studies rely on pairwise comparisons, where traits of each gene are compared with 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 PICs (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.

expression through the course of embryonic development for 10 animal species, each from a different phylum. Using pairwise comparisons, they found that there is more evolutionary variance in gene expression at a midphase 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, which is in contrast with the "hourglass" model previously proposed for closely related species (20). Furthermore, they suggested that this provides biological justification for the concept of phyla. We previously described concerns with the interpretations of this result (21). Here, we address the analyses themselves by examining the structure of the pairwise comparisons.

Results and Discussion

KMRR (16) Reanalysis.

Original pairwise test of the ortholog conjecture. KMRR (16) sought to test the ortholog conjecture. The ortholog conjecture (17) is the proposition that orthologs (genes that diverged from each other because of a speciation event) have more similar attributes than paralogs (genes that diverged from each other because of a gene duplication event). The ortholog conjecture can and has been applied to diverse attributes, including molecular sequence, biochemical function, and as in the study considered here, expression. It 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 homologous genes that have not been investigated in detail. 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 after 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 (22, 23). Analyses of domain structure were consistent with the ortholog conjecture (24). There have been few tests of the ortholog conjecture with regards to gene expression (22), and the study by KMRR (16) is one of the most detailed to date.

KMRR (16) considered several publicly available datasets of gene expression across tissues and species. Their expression summary statistic is tau (25, 26), an indicator of tissue specificity of gene expression. Tau can range from a value of zero, which indicates no specificity (i.e., uniform expression across tissues), to a value of one, 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, although of course, since the original expression is multidimensional, this means much information is discarded. This includes information about the tissue to which expression is specific. For example, if one gene has expression specific to the brain and another expression specific to the kidney, both would have a tau of one.

The analyses by KMRR (16) are based on pairwise comparisons (Fig. 1*A*) between tau within each gene family. Rather than make every pairwise comparison within each gene tree, they considered only a subset of pairwise comparisons in each particular analysis. They first selected a focal species, which varied from analysis to analysis. Ortholog comparisons were limited to pairs that include this species, and the only paralogs considered were those with the highest expression in this species. Note that this subset of pairwise comparisons still samples the same changes multiple times.

They found the correlation coefficient of tau for orthologs to be significantly greater than the correlation coefficient of tau for paralogs (i.e., orthologs tend to have more similar expression 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" (16).

Phylogenetic reanalyses. We reanalyzed the study by KMRR (16) using phylogenetic comparative methods. We focused on one of the datasets included in their analyses: that of Brawand et al. (27). This dataset is the best sampled in their analyses. It has gene expression data for six organs across 10 species (nine mammals and one bird), 8 of which were analyzed by KMRR (16) and are further considered here.

Many phylogenetic comparative methods are now available for addressing a wide range of questions relevant to comparative functional genomics. Some recent methods have been developed specifically for analyses of expression. For example, the Expression Variance and Evolution model compares the ratio of within-species expression variance with between-species evolutionary expression variance in an explicit phylogenetic framework (10). It can test a variety of hypotheses, including lineage-specific expression level shifts. This method considers strict orthologs that all have the same gene tree, which is the same as the phylogeny of the species under consideration. However, for the reanalysis presented here, we address different questions that consider a broad diversity of gene trees that include both speciation and different patterns of duplication. Phylogenetic independent contrasts (PICs) (2), the

most widely used phylogenetic comparative method, are particularly well-suited to this challenge.

For each internal node in each gene tree, we calculated the PIC (2) 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 (Fig. 1*B*). 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). This is similar to previous applications of contrasts to examine shifts in evolutionary rates between clades (28). Rather than compare the distributions of contrasts for nodes in different clades, as this previous work did, we compare the distributions of contrasts for nodes with different annotations (speciation or duplication).

We mapped the tau values calculated by KMRR (16) for the dataset by Brawand et al. (27) onto 21,124 gene trees parsed from ENSEMBL Compara (29). These are the same precomputed trees on which the orthology/paralogy annotations that KMRR (16) used are based; 8,854 gene trees passed taxon sampling criteria (four genes) after removing tips without tau values and had at least one speciation event. Of these, 8,516 were successfully time calibrated. These calibrated trees were used to calculate PICs for 21,017 duplication nodes and 67,845 speciation nodes (Tables S1 and S2). One of these trees is presented in Fig. S1 to show the 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 after speciation or duplication events (30). Under this hypothesis, we would predict that contrasts across speciation nodes and duplication nodes are drawn from the same distribution. Under the alternative hypothesis specified by the ortholog conjecture (that there is a higher rate of change after duplication events than speciation events), we would expect to see the distribution of duplication contrasts shifted to higher values relative to the speciation contrasts.

We did not find increased evolutionary change in expression after duplication events compared with speciation events (Fig. 2*B*). The Wilcoxon rank test does not reject the null hypothesis that the rate of evolution after duplications is the same as or less than the rate after speciation ($P = 1$). Our phylogenetic comparative analysis, unlike the previously published pairwise comparative analysis (16), therefore finds no support for the ortholog conjecture in this system. This result holds when we exclude the genes with the lowest phylogenetic signal according to Blomberg's K (31) (Fig. S2) and is robust to model selection (32–34) (*SI Text*).

We examined the possibility that ascertainment biases were differentially impacting the inference of expression evolution after 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 (Fig. S3). 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 (35, 36). 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 (*SI Text*).

An additional concern is that within-species expression variation is not considered when making comparisons of tau across homologous genes. Neglecting within-species variation is expected to mislead phylogenetic comparative analyses in some cases (10, 37, 38). Since branch length describes expected trait variation (2), within-species variation can be modeled by extending the terminal branches of each gene phylogeny (37). We modeled within-species variation by extending each terminal branch by 7 My. This value was selected, because it corresponds to the age of the clade Hominini (defined by the most recent common ancestor of humans and chimpanzees), the shallowest node in the phylogeny of examined species. This provides an opportunity to assess the impact that extreme within-species variation (i.e., as large as the variance between humans and chimpanzees and their most recent common ancestor) would have on our findings. These analyses still do not reject the null hypothesis and do not provide support for the ortholog conjecture (Fig. S4).

Understanding the incongruence between pairwise and phylogenetic methods. To better understand why our phylogenetic analysis supports a different conclusion (i.e., no support for the ortholog conjecture) than the published analysis by KMRR (16) (i.e., strong support for the ortholog conjecture), we first checked to make sure that we could reproduce their result based on pairwise analyses. This is important, since we are only looking at a subset of the data that they considered: the dataset by Brawand et al. (27) for gene trees that could be successfully time calibrated. In figure 1 in ref. 16, they present a higher tau correlation coefficient between ortholog pairs than between paralog pairs. We find the same here, with correlation coefficients of 0.75 ($P < 2 \times 10^{-16}$) for orthologs and 0.36 ($P < 2 \times 10^{-16}$) for paralogs.

Why is it that pairwise methods and phylogenetic methods lead to opposite conclusions? One reason 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 nonoverlapping branches of the tree. The other reason is that pairwise comparisons and phylogenetic comparative methods describe different things. Pairwise comparisons describe contemporary patterns, while phylogenetic methods infer historical processes (15). Paralogs can be more different from orthologs, even when the processes of evolution after speciation and duplication are the same. Any difference could be caused by 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 332.6 My, the mean distance between paralogs is 1,763.5 My. 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 (Fig. S3*A*) 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 (Fig. S3*B*), the paralog correlation coefficient increases from 0.36 to 0.55 ($P < 2 \times 10^{-16}$). This is much closer to the ortholog correlation of 0.75.

The study by KMRR (16) did investigate the impact of node age on correlation but in a different way. In figure 2 in ref. 16, 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 (Fig. 2*A*). There are, however, several difficulties with interpreting this plot. First, it does not just reflect the evolutionary processes that generated the data, but it is also impacted by the phylogenies along which these processes

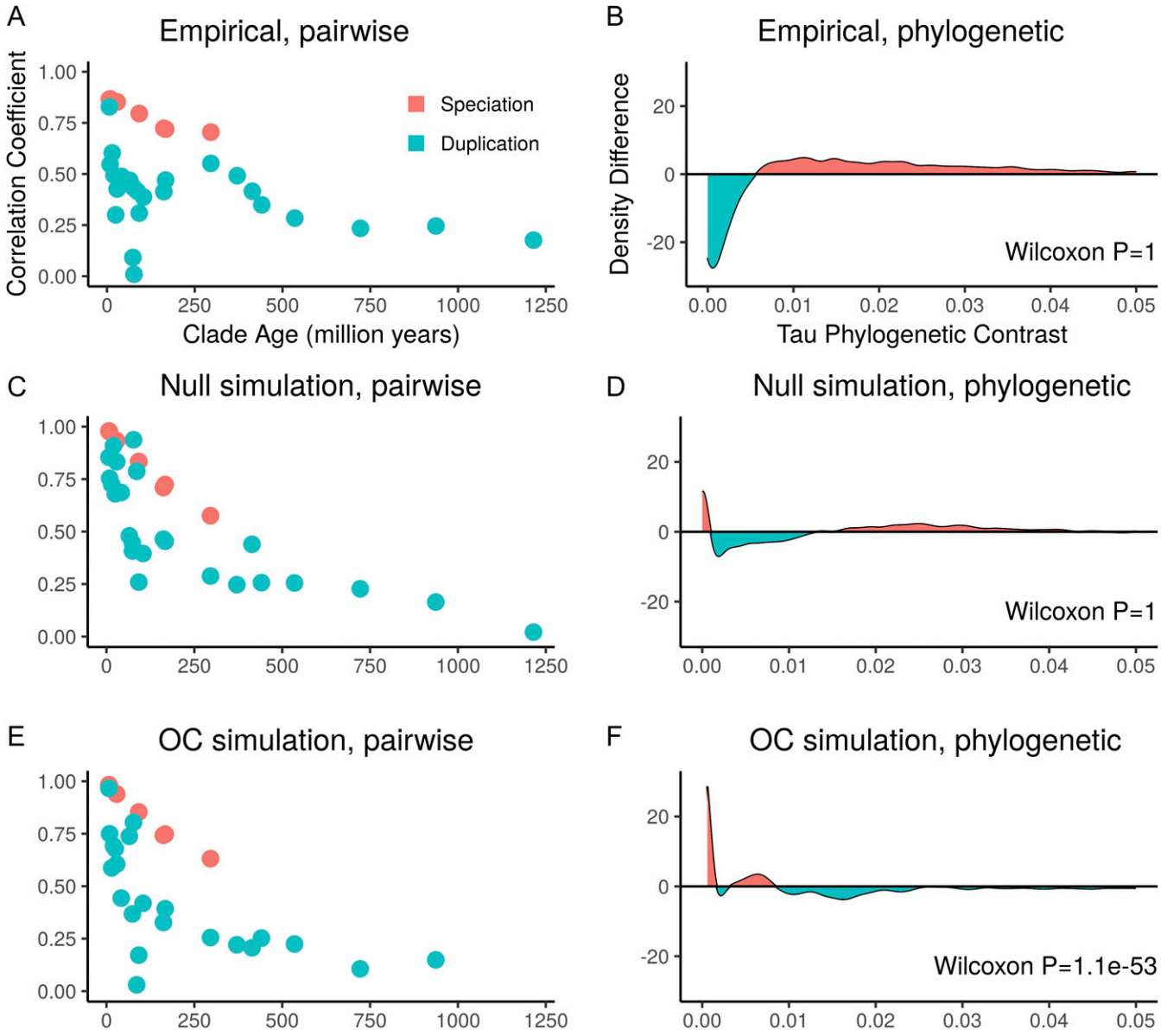

Fig. 2. Pairwise (*A*, *C*, and *E*) and phylogenetic (*B*, *D*, and *F*) analyses of the original data (*A* and *B*), data simulated under the null hypothesis (*C* and *D*), and data simulated under the ortholog conjecture (*E* and *F*). In the pairwise plots, each point indicates the correlation coefficient of tau for a set of pairwise comparisons annotated with a specific node name (e.g., Primates) and event type (speciation or duplication, giving rise to orthologs and paralogs, respectively). The phylogenetic plots show the difference between the density distributions for tau phylogenetic contrasts for speciation and duplication events, where a value above zero indicates an excess of speciation contrasts in the indicated interval. A horizontal line at zero would indicate that the density distributions are identical. *A* reproduces the pattern presented in figure 2A of the work by KMRR (16) of higher correlation across speciation events than duplication events, which they took as evidence of the ortholog conjecture. The recovery of a similar pattern under both simulations (*C* and *E*) indicates that this pairwise approach does not make distinct testable predictions. The phylogenetic analysis of the original data (*B*) does not show an excess of larger contrasts for duplication events and does not reject the null hypothesis, providing no support for the ortholog conjecture. *D* and *F* validate the phylogenetic approach by showing that it does not reject the null when data are simulated under the null (*D*) but does reject the null when data are simulated under the ortholog conjecture (*F*). OC, ortholog conjecture.

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 nonindependent pairwise comparisons. Third, instead of using the actual age of duplication events, the age of an adjacent named node in the species tree is used.

To better understand this plot (Fig. 2*A*), we performed simulations of tau on the calibrated gene trees. We did not modify the gene tree topologies or their inferred histories of duplication and loss. We simulated the evolution of tau under the null model that it evolves at the same rate after duplication and speciation events. Under the null model, the plot of correlation coefficient to node age (Fig. 2*C*) is very similar to that of the observed data (Fig. 2*A*). As in the original study, there is higher correlation coefficient across orthologs (0.73, $P < 2 \times 10^{-16}$) than paralogs (0.28, $P < 2 \times 10^{-16}$) when not considering node age. Phylogenetic analysis of the data simulated under the null hypothesis (Fig. 2*D*) does not reject the null hypothesis (Wilcoxon $P = 0.999$) as expected.

We next simulated the evolution of tau under the ortholog conjecture, where the rate of evolution of tau after duplication

was twofold the rate after speciation. The pairwise results of this heterogeneous model (Fig. 2*E*) are nearly indistinguishable from the results under the null model (Fig. 2*C*) and also, have a higher correlation coefficient for orthologs (0.76, $P < 2 \times 10^{-16}$) than paralogs (0.24, $P < 2 \times 10^{-16}$). The phylogenetic analysis of the ortholog conjecture simulation (Fig. 2*F*) does reject the null hypothesis (Wilcoxon $P = 1.1 \times 10^{-53}$) as expected if the phylogenetic methods are working correctly.

These simulations have several implications. The pairwise comparisons used by KMRR (16) cannot distinguish between the null hypothesis and ortholog conjecture. The pairwise results are strikingly similar under both hypotheses (Fig. 2 *C* and *E*). These simulations also serve to validate the phylogenetic methods applied to this problem. As expected, 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 can test explicit predictions based on hypotheses about evolutionary process.

Since greater rates of expression evolution after duplication do not explain the lower correlation of tau values in the pairwise comparison plots (Fig. 2 *A*, *C*, and *E*), this pattern must reflect some other property that is shared across these analyses. We found that the lower correlations for duplication comparisons can largely be explained by the greater variance in the age of duplication nodes. Each point in the pairwise correlation plot (Fig. 2 *A*, *C*, and *E*) summarizes multiple pairwise comparisons across nodes of a given event type (speciation or duplication as indicated by color) and clade in the species tree (Theria, Mammalia, Amniota, etc. as indicated by the age of the clade along the *x* axis). These clade names, in turn, are derived from the node annotations in the ENSEMBL Compara trees, which apply clade names to both speciation and duplication nodes. While speciation nodes in the gene trees have a clear correspondence to clades in the species trees, duplication nodes do not, since duplication can occur at any point along any branch in the species tree. Neighboring clade ages are, therefore, a very rough approximation of the age of duplication events. This is apparent when node ages derived from calibrated trees are plotted against the age of the clade annotations for each node (Fig. S5*A*). Because duplication events that are all annotated with the same clade name can occur at very different times, pairwise comparisons across these nodes capture evolutionary changes in tau across very different branch lengths. They, therefore, have lower correlation than pairwise comparisons made across speciation nodes. Of the 960,481 duplication nodes considered here, 264,300 have a node age (as determined by the time-calibrated trees) within 10% of their clade annotation. If only these duplication events closest to their annotated clade nodes are retained, the pattern of lower correlation of tau evolution for duplication events disappears for the dataset simulated under the null model (Fig. S5*B*). This correction, however, comes at the high cost of discarding 72.5% of duplication nodes.

Implications for the ortholog conjecture. There has been considerable recent interest in, and controversy about, the ortholog conjecture (17, 18, 30, 39, 40). While some studies have presented support for the ortholog conjecture, our results are consistent with multiple studies that have not (17, 30, 41). It should be noted that many areas of biology restrict the term "function" to describe the biochemical function of a gene, but the literature on gene duplication often uses function in a broader sense to include any gene attribute that could impact fitness, such as biochemical function, spatial expression, or binding. Since the ortholog conjecture can be applied to such diverse attributes, it is a heterogeneous and potentially idiosyncratic topic. For example, a recent integrated analysis of multiple gene attributes after duplication found extensive divergence in protein sequence but less divergence in expression (42). In our analyses, we found that phylogenetic distance is a better predictor of the similarity of gene expression than the history of gene duplication and speciation. The ortholog conjecture does not have to be an all or nothing proposition, even as it applies to expression. It may be that the rates of tissue-specific expression evolution after duplication are greater in some organisms, gene families, and evolutionary processes (39). The lack of support that we find for the ortholog conjecture in this system as well as the mixed results of others do suggest that investigators should test for it in each situation rather than assume a priori that it is a dominant pattern in the diversity of gene expression. These tests should be done in an explicit phylogenetic framework.

While empirical support for the ortholog conjecture has been mixed, theoretical work has suggested several mechanisms that could lead to increased rates of evolutionary change in expression and other gene traits after gene duplication. This has been a primary motivation for the extensive discussion of the ortholog conjecture. In his seminal work, Ohno (43) outlined three outcomes of gene duplication, now often referred to as neofunctionalization (one gene copy can take on new functions, while the other retains the old functions), subfunctionalization (each gene copy has a subset of the functions present in their most recent common ancestor), and conservation (each gene copy retains the same function). Neofunctionalization was the primary focus of early work on gene duplication, although it was not found to be a widespread outcome of duplication (44). The highly influential duplication–degeneration–complementation (DDC) model (45, 46) suggested that subfunctionalization could be a better explanation for the retention of both copies of a gene. It is built on the idea that gene copies that have each lost different functions would both be needed to fulfill all of the functions present in their most recent common ancestor. Tests of the DDC model have had mixed results (47). A growing body of work suggests that retention and differentiation of genes are caused by a variety of different mechanisms in different contexts and that we should not necessarily expect one process to dominate over others (48). Our lack of support for the ortholog conjecture suggests that whatever mechanisms lead to the retention of duplicated genes in the organisms considered here do not lead to changes in the tissue specificity of gene expression as summarized by tau.

The DDC model is particularly interesting to further consider in this context, since the summary statistic tau describes tissue specificity of expression. In the context of tissue-specific expression, the ortholog conjecture predicts a greater rate of evolutionary change in tau after duplication without specifying the direction of the change. The DDC model predicts a specific type of change in expression after duplication—subfunctionalization, which would increase tissue specificity, and therefore tau, after duplication. To test for this more specific prediction of DDC, we conducted additional analyses to see if there is a greater rate of increase in tau after duplication. For each branch in each gene tree, we inferred the change in tau and standardized this change by branch lengths. We found no evidence of increased tau after duplication relative to speciation (Wilcoxon $P = 0.846$). This result also holds when we discarded the 75% of genes with the lowest phylogenetic signal (Wilcoxon $P = 0.715$). These results are consistent with another recent study that found little support for neofunctionalization or subfunctionalization in expression of duplicated human genes and instead, proposed that dosage compensation drives most changes in expression after duplication in mammals (49). Like the ortholog conjecture, DDC is not an all or nothing hypothesis. We fail to find support for it in these tissues in these species with the tau expression summary statistic, but it may hold in other contexts. DDC could still play a role in the retention of these genes but with respect to other attributes, such as biochemical functions, rather than tissue-specific expression.

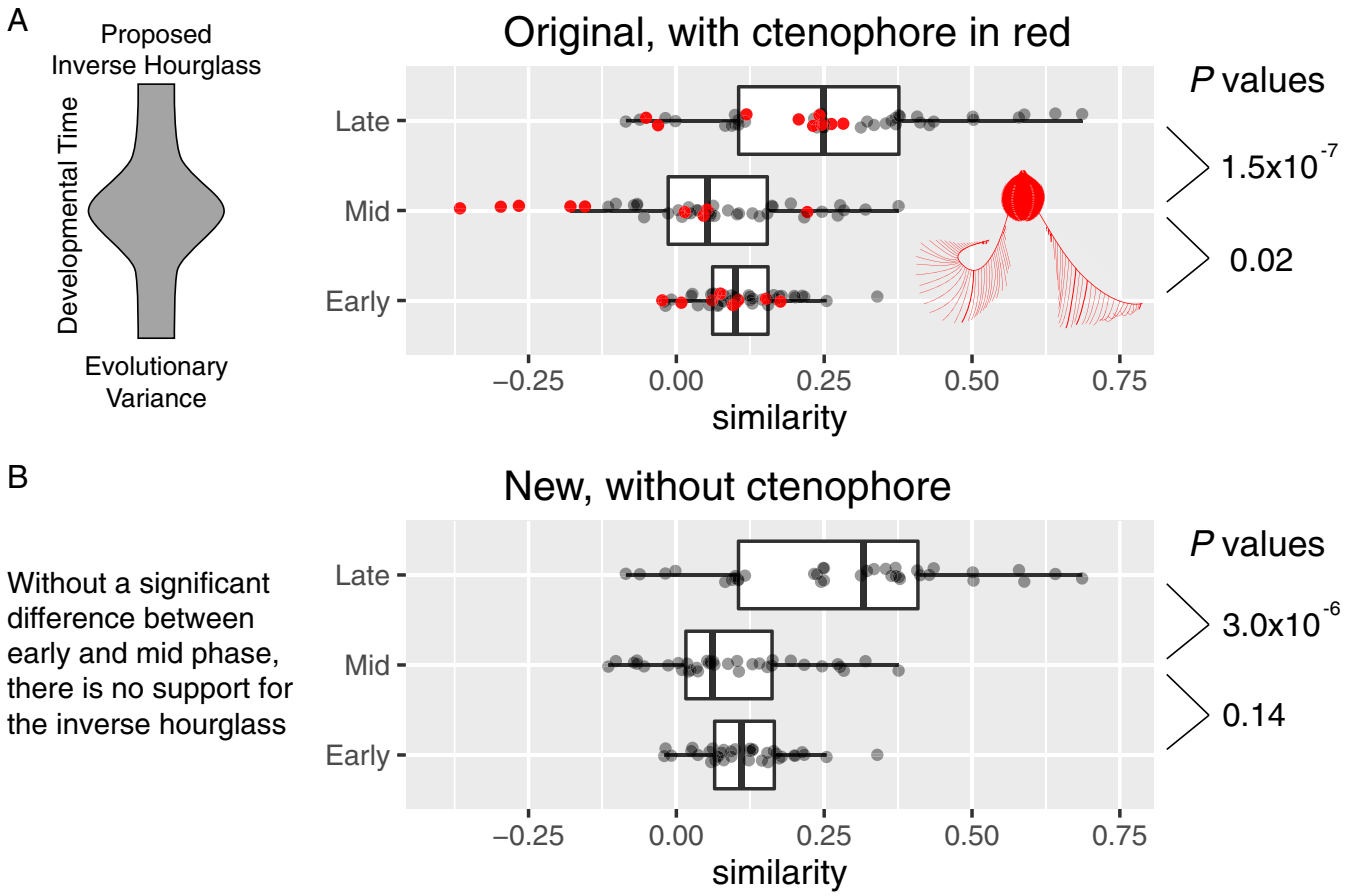

Fig. 3. Distributions of pairwise similarity scores for each phase of development. Pairwise scores for the ctenophore are red. Wilcoxon test *P* values for the significance of the differences between early–mid distributions and late–mid distributions are on the right. Model of variance, which is inversely related to similarity, is on the left. (*A*) The distributions as published by Levin et al. (19). Low similarity (i.e., high variance) in the midphase of development was interpreted as support for an inverse hourglass model for the evolution of gene expression. The five least similar midphase scores were all from the ctenophore. (*Inset*) The ctenophore image is by S. Haddock and reproduced from phylopic.org. (*B*) The distributions after the exclusion of the ctenophore. The early-phase and midphase distributions are not statistically distinct.

Levin et al. (19) Reanalysis.

Original pairwise analyses of developmental gene expression. Levin et al. (19) analyzed gene expression through the course of embryonic development for 10 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 middevelopmental 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 midphase 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 (20, 50). Levin et al. (19) 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. (19) 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 figure 4D in ref. 19, and their Kolmogorov–Smirnov (KS) tests indicated that the early distribution and late distribution were each significantly different from middistribution ($P < 10^{-6}$ and $P < 10^{-12}$, respectively). This is the support that they presented for the inverse hourglass model.

Pairwise comparisons oversample lineage-specific changes. We examined the matrix of pairwise comparisons used as the base for the KS tests and figure 4D in the work by Levin et al. (19) and thus, as support for 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 pairwise comparisons. We found that every data point was included twice, because both reciprocal pairwise comparisons (which have the same values) were retained. As a consequence, there are 90 entries for the 45 pairwise comparisons, and by doubling the data, the significance of the result seems stronger than it actually is. After removing the duplicate values, the *P* values are far less significant: 0.002 for the early to middle comparison and on the order of 10^{-6} for early to late. In addition, the test that they used (the KS test) is not appropriate

for the hypothesis that they seek to evaluate. The KS test does not just evaluate whether one distribution is greater than the other; it also tests whether the shapes of the distributions are the same. The samples in this dataset are matched (i.e., for each pairwise comparison, there are early, mid, and late expression values), which the KS test does not take into account. The paired version of the Wilcoxon test is instead appropriate in this case. When applied to the deduplicated data, the P value of this test is 0.02 for the early to middle comparison and on the order of 10^{-7} for early to late comparison.

After we addressed the issues above, 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 midphase distribution (Fig. 3*A*) 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 midphase distributions ($P = 0.14$ for the early to middle comparison and $P < 10^{-5}$ for the late to middle comparison) and no support for the inverse hourglass (Fig. 3*B*). This highlights a well-understood property of pairwise comparisons across species (2, 51): evolutionary changes along a given branch, like those along the ctenophore branch, impact each of the multiple pairwise comparisons that include that branch. The pairwise comparisons are, therefore, not independent—different pairwise comparisons are impacted by changes along some of the same branches (Fig. 1*A*). 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 show problems with pairwise comparisons that impact the analysis by Levin et al. (19), we did not perform a phylogenetic reanalysis of this study as we did for the study by KMRR (16). This is because the similarity metric computed in the pairwise comparisons of Levin et al. (19) is not suitable for phylogenetic analysis, as it is based on different genes for different species pairs, and therefore, it cannot be modeled across the phylogeny. A full phylogenetic reanalysis would be possible using upstream analysis products to rederive new expression summary statistics.

Phylogenetic comparative methods in functional genomics. We are not the first to apply phylogenetic comparative methods to functional genomic data. While the vast majority of comparative functional genomic studies have used standard pairwise similarity methods, a small number of comparative functional genomic studies have used phylogenetic comparative approaches (52–55). For instance, a phylogenetic ANOVA (10) of the evolution of gene expression in strict orthologs improves statistical power and drastically reduces the rate of false positives relative to pairwise approaches.

Some of the most widely used phylogenetic comparative methods (2, 3) are already directly applicable to comparative functional genomic studies. There are also interesting new challenges, such as parameterizing differential expression, so that species-specific technical biases are not mistaken for evolutionary changes in expression (56).

Branch lengths are fundamental to phylogenetic comparative methods, since the null expectation is that there are more changes along longer branches. In this study, we applied four separate modifications of branch lengths. First, we time calibrated the gene trees, so that nodes corresponding to the same speciation events had the same age across all trees. This is critical to making the comparisons of functional genomic traits equivalent across genes. Second, PICs (including the implementation that we use here) adjust internal branch lengths to accommodate the variation associated with estimating character states at internal nodes (2). Third, we randomized the calibration times to determine the sensitivity of the analyses to calibration times and the branch length estimations that they impact (*SI Text*). Fourth, we extended terminal branch lengths to understand the potential impact of within-species variation (Fig. S4). The first two of these modifications are technical requirements for the application of phylogenetic comparative methods. The second two explore the sensitivity of the analyses to branch length and the information that they convey and find that the results are robust to variation in branch length. This is encouraging for the future of phylogenetic comparative functional genomics given that one concern of applying these methods is the challenge of estimating branch lengths.

Conclusions

The problems that we identify with pairwise comparisons in two recent functional genomics studies indicate that there are likely to be similar problems in other studies that use these methods. Future studies that compare functional genomic data across species 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.

These analyses also illustrate how important it is to not conflate evolutionary patterns with the processes that generated them. Finding a pattern where paralogs tend to be more different from 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. 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 genomic traits across species as they are to comparing morphology or any of the other traits to which they are already routinely applied.

Methods

All files needed to reexecute the analyses presented in this document are available at https://github.com/caseywdunn/comparative_expression_2017. The most recent commit at the time that the analyses were executed was b9601ebb. The most recent commit at the time that the manuscript was rendered was 242e2146.

KMRR (16) Reanalysis. The study by KMRR (16) followed excellent practices in reproducibility. They posted all data and code needed to reexecute 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 obtained the Compara.75.protein.nh.emf gene trees (29) from ftp://ftp.ensembl.org/pub/release-75/emf/ensembl-compara/homologies/ and include them in our github repository. These gene trees include branch lengths, annotate each internal node as being a duplication or speciation event, and provide a clade label for each internal node.

We considered only the data from the work by Brawand et al. (27) for the eight taxa included in the work by KMRR (16). We left in sex chromosome

genes and testes expression data, which KMRR (16) removed in some of their sensitivity analyses. This corresponded to the analyses by KMRR (16) 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 (57), we added tau estimates generated by the KMRR (16) Rscript.R to the tree data objects. We then pruned away tips without expression data, retaining only the trees with four 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 gene trees. This, in turn, scales the PICs (which take branch length into account) so that they can be compared appropriately. Before calibrating the trees, we had to slightly modify some of them. The node names in the ENSEMBL Compara (29) gene phylogenies are parsed from the National Center for Biotechnology Information (NCBI) Taxonomy database, which has many polytomies, rather than a bifurcating species phylogeny. One implication of this is that node names can be resolved so that a speciation node can have the same name as one of its speciation node ancestors, as others have noted (58). If left unaddressed, this would force all intervening branches to have length zero and interfere with calibration. In particular, Hominini is the name for the clade that includes humans and chimps, while Homininae is the clade that includes humans, chimps, and gorillas. Because of the structure of the NCBI Taxonomy, both clades are labeled as Homininae in the Compara trees. To remedy this, 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 (16) code, with the exception of Hominini and Homininae. These we set to 7 and 9 My, respectively, drawing on the same TimeTree source (59) that KMRR (16) used. We used the chronos() function from the R package ape (60) for this calibration with the correlated model. *SI Text* has 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 PIC for tau across its daughter branches with the pic() function in ape (60). Summary statistics on taxon and node sampling are presented in Tables S1 and S2. We then collected the contrasts from all trees into a single data frame along with other annotations, including whether the node is a speciation or duplication event. This data frame, nodes_contrast, was then analyzed as described in the text for the presented plots and tests.

Levin et al. (19) Reanalysis. Levin et al. (19) helpfully provided data and clarification on methods. We obtained the matrix of pairwise scores that underlies their figure 4D (19) and confirmed that we could reproduce their published results. We then removed duplicate rows, applied the Wilcoxon test in place of the KS test, and identified ctenophores as overrepresented among the low outliers in the middevelopmental 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. We thank Steve Haddock, Alex Damian Serrano, August Guang, Bruno Vellutini, Zack Lewis, and Matthew Hahn for feedback on the manuscript. Marc Robinson-Rechavi provided information about the KMRR (16) analyses and suggestions for our manuscript. The Ensembl Comparative Genomics team, including Matthieu Muffato, provided support for helping us better understand and use their Compara trees. The contribution of C.W.D. was supported by National Science Foundation Grant DEB-1256695 and the National Science Foundation Waterman Award. A.H. was supported by the European Research Council Community's Framework Program Horizon 2020 (2014–2020) European Research Council Grant 648861.

1. Wray GA (2013) Genomics and the evolution of phenotypic traits. *Annu Rev Ecol Evol Syst* 44:51–72.
2. Felsenstein J (1985) Phylogenies and the comparative method. *Am Nat* 125:1–15.
3. Grafen A (1989) The phylogenetic regression. *Philos Trans R Soc Lond B Biol Sci* 326: 119–157.
4. Pagel M (1999) Inferring the historical patterns of biological evolution. *Nature* 401: 877–884.
5. Revell LJ, Mahler DL, Peres-Neto PR, Redelings BD (2012) A new phylogenetic method for identifying exceptional phenotypic diversification. *Evolution* 66:135–146.
6. FitzJohn RG (2012) Diversitree: Comparative phylogenetic analyses of diversification in R. *Methods Ecol Evol* 3:1084–1092.
7. Uyeda JC, Harmon LJ (2014) A novel Bayesian method for inferring and interpreting the dynamics of adaptive landscapes from phylogenetic comparative data. *Syst Biol* 63:902–918.
8. Garamszegi LZ (2014) *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology* (Springer, Berlin).
9. Jhwueng D-C (2013) Assessing the goodness of fit of phylogenetic comparative methods: A meta-analysis and simulation study. *PLoS One* 8:e67001.
10. Rohlfs RV, Nielsen R (2015) Phylogenetic ANOVA: The expression variance and evolution model for quantitative trait evolution. *Syst Biol* 64:695–708.
11. Barker D, Pagel M (2005) Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. *PLoS Comput Biol* 1:e3–8.
12. Garland T, Jr, Bennett AF, Rezende EL (2005) Phylogenetic approaches in comparative physiology. *J Exp Biol* 208:3015–3035.
13. Ricklefs RE, Starck JM (1996) Applications of phylogenetically independent contrasts: A mixed progress report. *Oikos* 77:167–172.
14. Chamberlain SA, et al. (2012) Does phylogeny matter? Assessing the impact of phylogenetic information in ecological meta-analysis. *Ecol Lett* 15:627–636.
15. O'Meara BC (2012) Evolutionary inferences from phylogenies: A review of methods. *Annu Rev Ecol Evol Syst* 43:267–285.
16. Kryuchkova-Mostacci N, Robinson-Rechavi M (2016) Tissue-specificity of gene expression diverges slowly between orthologs, and rapidly between paralogs. *PLOS Comput Biol* 12:e1005274–13.
17. Nehrt NL, Clark WT, Radivojac P, Hahn MW (2011) Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLOS Comput Biol* 7: e1002073.
18. Koonin EV (2005) Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* 39:309–338.
19. Levin M, et al. (2016) The mid-developmental transition and the evolution of animal body plans. *Nature* 531:637–641.
20. Kalinka AT, et al. (2010) Gene expression divergence recapitulates the developmental hourglass model. *Nature* 468:811–814.
21. Hejnol A, Dunn CW (2016) Animal evolution: Are phyla real? *Curr Biol* 26:R424–R426.
22. Chen X, Zhang J (2012) The ortholog conjecture is untestable by the current gene ontology but is supported by RNA sequencing data. *PLOS Comput Biol* 8:e1002784.
23. Thomas PD, Wood V, Mungall CJ, Lewis SE, Blake JA; Gene Ontology Consortium (2012) On the use of gene ontology annotations to assess functional similarity among orthologs and paralogs: A short report. *PLOS Comput Biol* 8:e1002386.
24. Forslund K, Pekkari I, Sonnhammer EL (2011) Domain architecture conservation in orthologs. *BMC Bioinformatics* 12:326.
25. Yanai I, et al. (2005) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* 21:650–659.
26. Kryuchkova-Mostacci N, Robinson-Rechavi M (2016) A benchmark of gene expression tissue-specificity metrics. *Brief Bioinform* 18:205–214.
27. Brawand D, et al. (2011) The evolution of gene expression levels in mammalian organs. *Nature* 478:343–348.
28. Garland T, Jr (1992) Rate tests for phenotypic evolution using phylogenetically independent contrasts. *Am Nat* 140:509–519.
29. Herrero J, et al. (2016) Ensembl comparative genomics resources. *Database (Oxford)* 2016:bav096–bav17.
30. Studer RA, Robinson-Rechavi M (2009) How confident can we be that orthologs are similar, but paralogs differ? *Trends Genet* 25:210–216.
31. Blomberg SP, Garland T, Jr, Ives AR (2003) Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. *Evolution* 57:717–745.
32. Akaike H (1974) A new look at the statistical model identification. *IEEE Trans Automat Contr* 19:716–723.
33. Posada D, Buckley TR (2004) Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Syst Biol* 53:793–808.
34. Pennell MW, et al. (2014) Geiger v2.0: An expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics* 30:2216–2218.
35. Takahata N (1989) Gene genealogy in three related populations: Consistency probability between gene and population trees. *Genetics* 122:957–966.
36. Degnan JH, Rosenberg NA (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol Evol* 24:332–340.
37. Ives AR, Midford PE, Garland T, Jr (2007) Within-species variation and measurement error in phylogenetic comparative methods. *Syst Biol* 56:252–270.
38. Felsenstein J (2008) Comparative methods with sampling error and within-species variation: Contrasts revisited and revised. *Am Nat* 171:713–725.
39. Gabaldón T, Koonin EV (2013) Functional and evolutionary implications of gene orthology. *Nat Rev Genet* 14:360–366.
40. Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C (2012) Resolving the ortholog conjecture: Orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLOS Comput Biol* 8:e1002514.
41. Yanai I, Graur D, Ophir R (2004) Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. *OMICS* 8:15–24.
42. Soria PS, McGary KL, Rokas A (2014) Functional divergence for every paralog. *Mol Biol Evol* 31:984–992.
43. Ohno S (1970) *Evolution by Gene Duplication* (Springer, Berlin).
44. Hughes AL (1994) The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci* 256:119–124.

45. Force A, et al. (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151:1531–1545.
46. Lynch M, Force A (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154:459–473.
47. Huminiecki L, Wolfe KH (2004) Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome Res* 14: 1870–1879.
48. Hahn MW (2009) Distinguishing among evolutionary models for the maintenance of gene duplicates. *J Hered* 100:605–617.
49. Lan X, Pritchard JK (2016) Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. *Science* 352:1009–1013.
50. Domazet-Lošo T, Tautz D (2010) A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* 468:815–818.
51. Ackerly DD, Reich PB (1999) Convergence and correlations among leaf size and function in seed plants: A comparative test using independent contrasts. *Am J Bot* 86:1272–1281.
52. Oakley TH, Gu Z, Abouheif E, Patel NH, Li W-H (2005) Comparative methods for the analysis of gene-expression evolution: An example using yeast functional genomic data. *Mol Biol Evol* 22:40–50.
53. Eng KH, Bravo HC, Keleş S (2009) A phylogenetic mixture model for the evolution of gene expression. *Mol Biol Evol* 26:2363–2372.
54. Chang D, Duda TF, Jr (2014) Application of community phylogenetic approaches to understand gene expression: Differential exploration of venom gene space in predatory marine gastropods. *BMC Evol Biol* 14:123.
55. Whitney KD, Boussau B, Baack EJ, Garland T, Jr (2011) Drift and genome complexity revisited. *PLoS Genet* 7:e1002092–5.
56. Dunn CW, Luo X, Wu Z (2013) Phylogenetic analysis of gene expression. *Integr Comp Biol* 53:847–856.
57. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y (2016) ggtree: An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* 8:28–36.
58. Daub J, Moretti S, Davydov II, Excoffier L, Robinson-Rechavi M (2016) Detection of pathways affected by positive selection in primate lineages ancestral to humans. bioRxiv:10.1101/044941.
59. Hedges SB, Dudley J, Kumar S (2006) TimeTree: A public knowledge-base of divergence times among organisms. *Bioinformatics* 22:2971–2972.
60. Paradis E, Claude J, Strimmer K (2004) APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* 20:289–290.
61. Hahn MW (2007) Bias in phylogenetic tree reconciliation methods: Implications for vertebrate genome evolution. *Genome Biol* 8:R141.