

Phylogenetic comparative methods are problematic when applied to gene trees with speciation and duplication nodes: correcting for biases in testing the ortholog conjecture

Tina Begum^{1,2}, Marc Robinson-Rechavi^{1,2*}

¹Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland

²SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland

*Corresponding author

E-mail: marc.robinson-rechavi@unil.ch

Abstract

Most comparative studies of functional genomics have used pairwise comparisons. Yet it has been shown that this can provide biased results, since genes, like species, are phylogenetically related. Phylogenetic comparative methods should allow to correct for this, but they depend on strong assumptions, including unbiased tree estimates relative to the hypothesis being tested. Such methods have recently been used to test the “Ortholog Conjecture”, the hypothesis that functional evolution is faster in paralogs than orthologs. Whereas pairwise comparisons of tissue specificity index (τ) provided support for the ortholog conjecture, it was not supported using phylogenetic independent contrasts. We find that the gene trees used suffer from important biases, due to the inclusion of trees with no duplication nodes, to the relative age of speciations and duplications, to systematic differences in branch lengths, and to non-Brownian motion of tissue-specificity on many trees. We find some support for the ortholog conjecture, but especially that incorrect implementation of phylogenetic method in empirical gene trees with duplications can be problematic.

Author Summary

There is a lot of interest in understanding the evolution of gene function. Comparing functional genomics results offers a way to do this. In most cases, it is done by comparing pairs of genes, and notably pairs of orthologs (homologs derived by speciation) and pairs of paralogs (homologs derived by duplication). A drawback of such pairwise comparisons is that they neglect the evolutionary history of the genes, and thus violate the statistical presumption of independence of observations. Phylogenetic comparative methods have been suggested to correct for this. Recent results indicate that whereas pairwise comparisons support stronger functional divergence of paralogs than of orthologs, phylogenetic analyses would not. Re-analyses of these data show that gene trees are biased in such a way as to cause biases in phylogenetic methods when there are gene duplication. This can lead to erroneous results, and have serious consequences in biological data interpretation.

Introduction

The “Ortholog Conjecture”, a cornerstone of phylogenomics, has become a topic of debate in recent years [1–9]. Dealing with the roles of gene duplications in functional evolution, the ortholog conjecture is routinely used by both experimental and computational biologists in predicting, or understanding gene function. According to this model, orthologs (i.e. homologous genes which diverged by a speciation event) retain equivalent or very similar functions, whereas paralogs (i.e. homologous genes which diverged by a duplication event) share less similar functions [1]. This is linked to the hypothesis that paralogs evolve more rapidly. This hypothesis was challenged by results suggesting that paralogs would be functionally more similar than orthologs [2]. Such findings not only raised questions on the evolutionary role of gene duplication but also questioned the reliability of using orthologs to annotate unknown gene functions in different species [10]. Several more recent studies [3–6,8] found support for ortholog conjecture, mostly based on comparisons of gene expression data.

While all previous studies of the ortholog conjecture had used pairwise comparisons of orthologs and paralogs, a recent article suggested that this was flawed, and that phylogenetic comparative methods should be used [7]. In the field of comparative biology, most of the studies still rely on the pairwise method including the testing of the ortholog conjecture. Ignorance of phylogenetic structure has been reported to underestimate the fundamental assumption of independent observations in statistics [11,12]. A solution is to use phylogeny-based methods to investigate such evolutionary patterns [11–17]. There are three main such phylogenetic methods: Phylogenetic Independent Contrast (PIC) [11], Phylogenetic Generalized Least-Square (PGLS) [16], and Monte Carlo computer simulation [17]. PIC is widely adopted for its relative simplicity, and its applicability to a wide range of statistical procedures [7,18]. The performance of PIC relies on three basic assumptions: a correct tree topology; accurate branch lengths; and trait evolution following a Brownian model (where trait variance accrues as a linear function of time) [12–15,18–20]. If any of these assumptions is incorrect, this can lead to incorrect interpretation of results. This is probably why the application of such phylogenetic methods is still limited, and debated even after being introduced about three decades ago [12,13,15,18,19]. Dunn et al. [7] took an innovative approach in applying PIC to compare the divergence rates between two different events (“speciation” and “duplication”) to test the ortholog conjecture. However, such an application might be problematic since the time of occurrence of gene duplication, one of the two types of events compared, is unknowable by external information (e.g. no fossil evidence). Therefore, further study is required to understand why Dunn et al. [7] obtained results which are inconsistent with most studies by using the phylogenetic method. It is possible that all the conclusions drawn by previous studies on gene duplication are incorrect due to overlooking phylogenetic tree structure. If so, it should be well supported.

We re-examined the data of Dunn et al., after reproducing it using the resources and scripts provided by the authors [7]. Our reanalysis highlights potential problems associated with phylogenetic independent contrasts when applied to the impact of gene duplication. Finally, with proper controls, the phylogenetic method supports the ortholog conjecture.

Results

Issues with naive use of Phylogenetic Independent Contrasts (PICs)

To understand their results, we first reanalyzed the data of Dunn et al. [7]. We were able to reproduce all the results published in their article by running the code, which they clearly supplied. Dunn et al. reported a non-significant result for the PIC under the null simulation, using a Wilcoxon one-tailed rank test to check if the contrasts of duplication events are higher than the contrasts of speciation events ($P = 1$). Surprisingly, the PIC rejects the null hypothesis on the null simulations with a Wilcoxon two-tailed rank test (Fig 1A), with significant support for higher contrasts after speciation than duplication. This was robust to repeating the simulations with different random seed number (S1 Fig). This indicates that neither of the methods, PIC or pairwise, worked properly for these calibrated trees, since both rejects the null when simulations are performed under null.

PIC is standardized by the expected variance of daughter branch lengths for each node of event [12–15]. Therefore, accurate branch length estimation in absolute time (e.g. in Million Years – My) is needed [12–15,18].

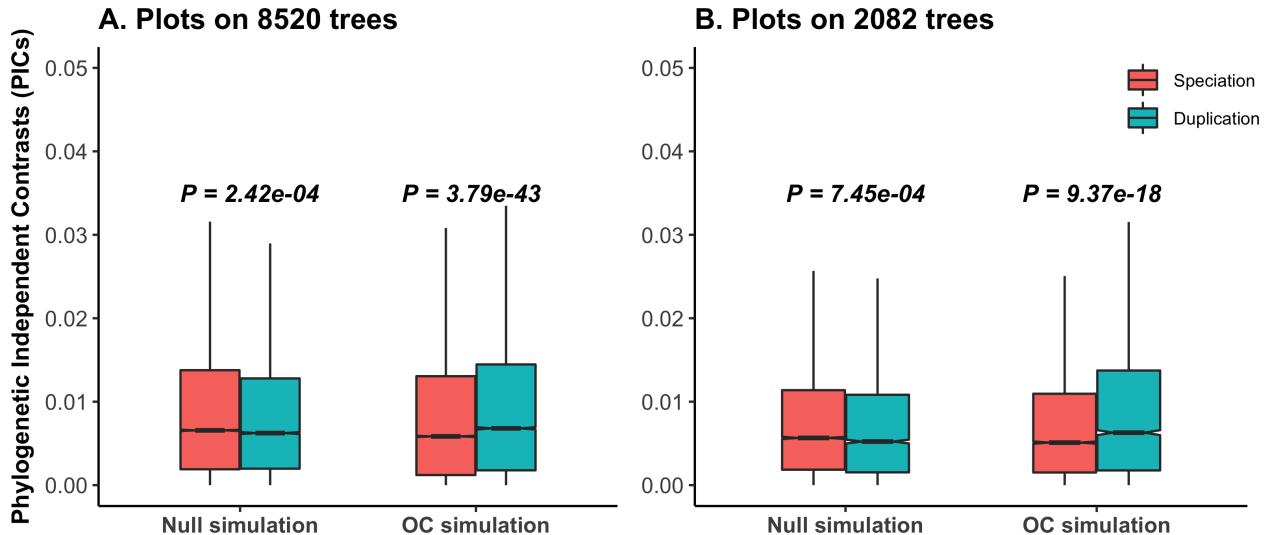


Fig 1: Reanalyses of phylogenetic simulation data of Dunn et al. [7]. P values are from Wilcoxon two-tailed tests. In null simulations, there should be no difference in contrasts between events. In OC (Ortholog Conjecture) simulations, contrasts are expected to be higher for duplication than for speciation. (A) An excess of contrasts for speciation than duplication reject the null hypothesis under null simulation scenario for all empirical time calibrated gene trees. (B) Results are similar with a subset of trees with strong phylogenetic signal for τ .

Since the accuracy and performance of the PIC method largely depend on proper branch length calibration [13,14,21] we investigated possible biases created during calibration of gene trees, especially concerning duplication times for which we have no external reference (e.g. no fossils) (S2 Fig).

The purpose of applying a phylogenetic method is to generate independent statistical data points for traits, which are otherwise not independent enough because of their phylogenetic relatedness. Such statistical non-independence among species trait values can be measured by “phylogenetic signal” [21–25]. Blomberg’s K and Pagel’s λ are widely used methods to capture such phylogenetic signal [21,22]. Dunn et al. [7] used Blomberg’s K in their study. In case of Blomberg’s K the value ranges from 0 to ∞ , where a value of 0 indicates no phylogenetic signal, a value of 1 indicates trait evolution according to a Brownian model (BM), values between 0 and 1 indicate that distant relatives are more similar than expected under BM, and values larger than 1 indicate that close relatives are more similar than expected under BM [21–25]. With a cutoff of $K > 0.551$, Dunn et al. [7] obtained only 2082 out of 8520 calibrated trees with strong phylogenetic signal. Using a cut-off of $P < 0.05$ together with $K > 0.551$ leads to 1135 trees, which produce similar results (S3 Fig) to that set of 2082 trees. We continued analyses with the 2082 trees of Dunn et al. [7] for consistency. In any case, it is notable that trait (τ) values are independent of phylogeny for the majority of the gene trees. For these 6438 trees, implementation of PIC to produce statistically independent data points might not be necessary. Moreover, it can be misleading if the contrasts are not checked for adequate standardization as per BM after applying PIC (discussed later). The phylogenetic method still rejects the null hypothesis under null simulations for those 2082 trees (Fig 1B), showing that the problem is not simply due to low phylogenetic signal.

Dunn et al. [7] used 7 speciation time points to calibrate gene trees. The oldest speciation calibration was at 296 My. All other calibrated nodes were duplication nodes leading to obtain ~652 unique duplication time points for the 8520 calibrated trees [7]. Out of these, 359 time points preceded the oldest speciation node age. Surprisingly, the calibrated node age of the oldest duplication event was found to be 11799977 My, that is, 2600 times older than the Earth. This is indicative of the difficulty of estimating the age of ancient duplications by phylogenetic methods. Those 359 high duplication node ages eventually led to much larger expected variances for gene duplication events (median expected variance for duplication events preceding the oldest speciation events: 828, median expected variance for speciation events: 184). This explains why

the mean paralog distance was ~5.6 times higher than that of the mean ortholog distance (S4A Fig). This data distribution also makes it appear as if no speciation happened before 296 My, a problem shared with the pairwise analysis of the same data [6]. There are also branch length issues for duplication events younger than the oldest speciation events, including very short branches and negative branch lengths. All this led to obtain abnormally high duplication contrasts for the empirical data. To limit such problems, Dunn et al. [7] removed node contrasts higher than 0.5 on the empirical data (this does not impact Fig 1, as the simulation data never has such high contrasts). Following this practice, there are still higher expected variances following gene duplication events (S4A and S4B Figs). In the null simulations only the τ values are simulated, while the branch lengths (hence the variances) are taken from the empirical data, and thus share its biases. This explains why contrasts are lower for duplications than for speciations under null simulations as well as with empirical data.

Randomization tests to assess the performance of phylogenetic method

We used randomization tests to assess whether the results of different analyses of the empirical dataset are reliable and unbiased. In a first randomization test, we randomized the trait values (i.e. randomly permuting the τ values) across the tips of each tree without altering the node events of the trees. We then computed the contrasts for the speciation and duplication events of the trees when there is no relation between trait values and phylogenetic relationships. When we compared the nodes contrasts of speciation and duplication events of these randomized trees (Fig 2A), we found the same pattern as reported for the empirical gene trees by Dunn et al. [7]. This shows that the PIC applied to these data is not measuring phenotypic evolution. It confirms that results are driven by their large differences in branch lengths (i.e. in expected variances) (Fig 2B), as on simulated null data. Any effect of trait divergence rates of speciation and duplication events is masked by this branch length difference. This violates the basic assumption of applicability of the PIC method to Brownian trait evolution. To remove the problem of difference in expected variances of the two events, we performed a second randomization test: we kept the original τ value for tips but randomly shuffled the events (duplication or speciation) of internal nodes of the empirical gene trees to maintain the original proportions of speciation and duplication events. The resulting trend (Fig 2C) still resembled the empirical gene trees data of Dunn et al. [7]. This appears due to the fact that the majority of the tree events are speciation events (Fig 2D) with speciation node ages ≤ 296 My. Most of the trees with many duplication events on the other hand have ancient duplication events (duplication node age > 296 My) for which the evolutionary rates of duplication are often masked by the effect of longer branch lengths. Hence, opposite to our expectation, the calibrated trees with no or few duplications have higher overall nodes contrast (apparent fast evolution) than trees with many duplications (apparent slow evolution). This might be due to greater difficulty in detecting paralogs for fast evolving genes. Therefore, reshuffling of the events may not change the observed pattern of higher speciation contrasts than duplication contrasts.

Out of 8520 calibrated trees, 2990 were species-only trees with no duplication events. For these 2990 trees, random shuffling of events had no impact. To avoid this bias, we removed those 2990 speciation trees as well as trees with negative branch lengths, and randomized the trait or the internal node events 100 times on the remaining 5479 trees. However, we still always obtained significantly higher contrasts of speciation than of duplication (S5A and S5B Figs). The randomization tests pattern is the same when we used 2082 trees with strong phylogenetic signals (S5C and S5D Figs). All these analyses indicate that the results reported by Dunn et al. [7] are biased by the available gene trees, and that this bias is not easy to correct.

The ortholog conjecture when τ evolution follows a Brownian model

Following a Brownian model (BM), the small and large values of standardized contrasts should be equally as likely to occur on any node of the tree [19]. Diagnostic plot tests (details in the Methods) for each tree can indicate whether trait evolution follows BM for that tree [13–15,18–20]. Since BM is intrinsic to the implementation of PIC method, we identified trees that passed diagnostic plot tests at 95% confidence level for τ . Hence, we used a subset of 2545 gene trees by removing trees with negative branch lengths, pure speciation trees, and trees that failed diagnostic tests for BM (S6 Fig). Moreover, to meet the correct branch length assumption of PIC applicability on trees with gene duplications, we used limitations of node age on gene duplication. It is necessary to put such a limitation to avoid biases due to inaccurate calibration of very old duplication nodes. Analyses on 2545 these trees with a node age limit of 296 My (i.e. node age \leq

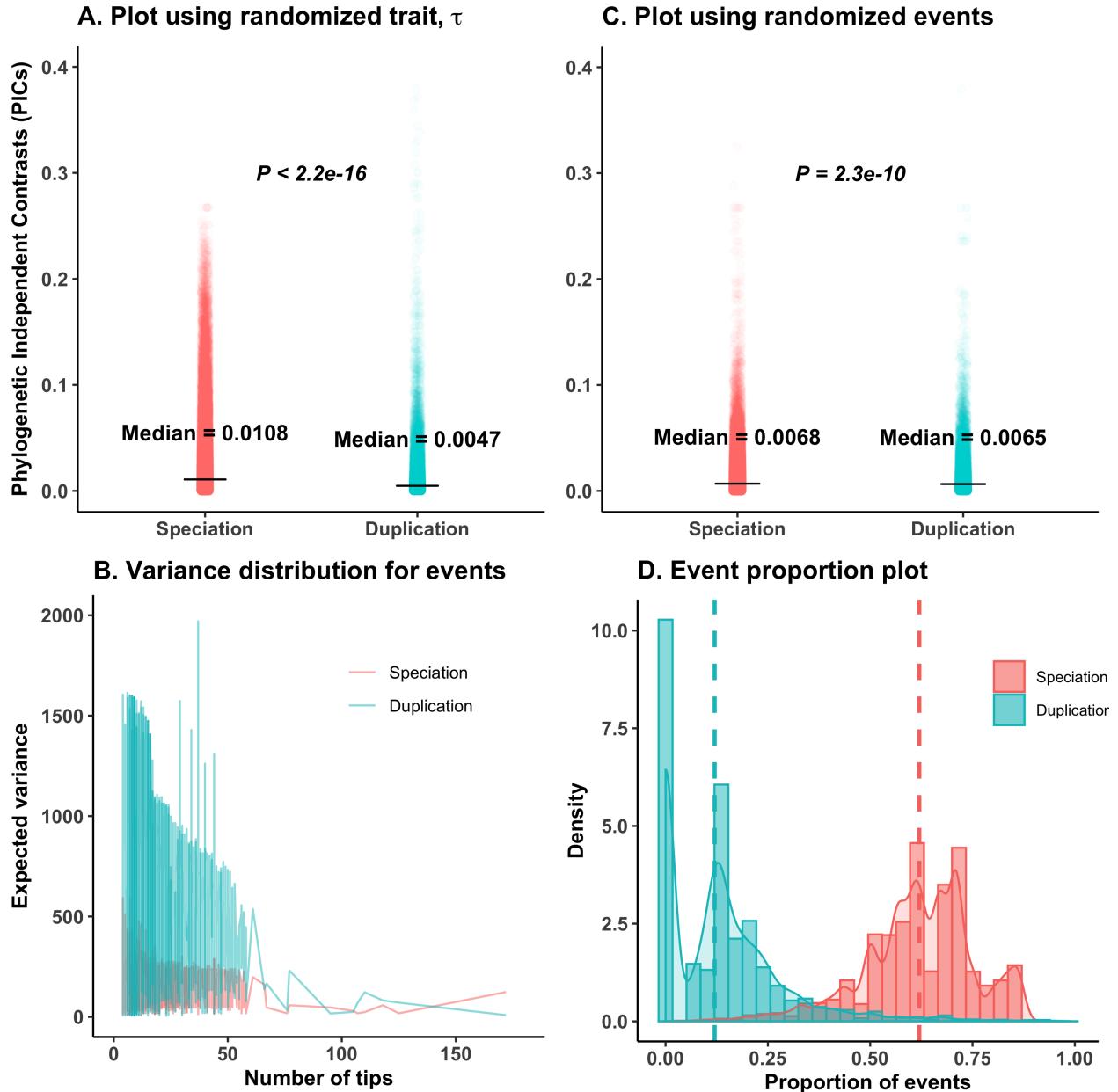


Fig 2: Analyses on calibrated empirical gene trees of Dunn et al. [7]. P values are from Wilcoxon two-tailed tests. (A) Randomly shuffling the τ values of the tips for 8520 gene trees does not alter the trend of empirical result [7] showing an opposite trend to the ortholog conjecture. (B) The figure indicates the expected variance is much higher for duplication than speciation events irrespective of the number of tips considered for the study. (C) Using the original τ data, if we permute the events (Speciation/Duplication/NA) of the nodes, the trend of result still resembles the empirical result obtained by the recent study [7]. (D) The proportions of speciation events are much higher than duplication events for all time-calibrated trees. The dotted line represents the median proportion of both events. This plot shows a high frequency of trees have no/few duplication events.

Plots on contrasts standardized 2545 gene trees

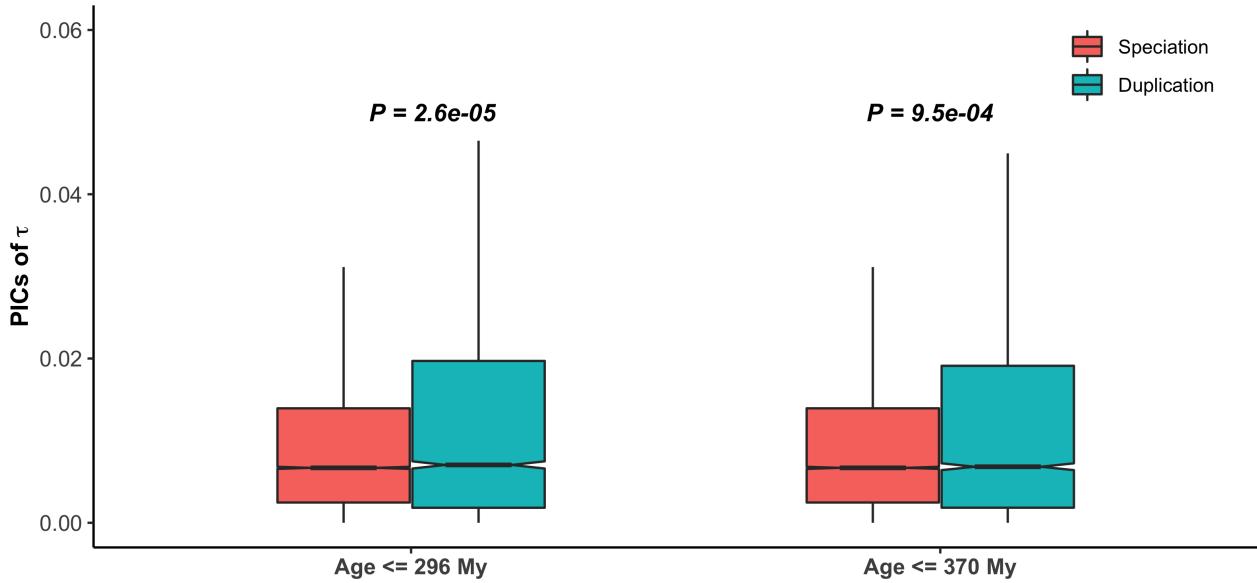


Fig 3: The ortholog conjecture test on τ for trees passing diagnostic plot tests with proper controls. P values are from Wilcoxon two-tailed tests. ‘PICs’ refers ‘Phylogenetic Independent Contrasts’. Contrasts of speciation and duplication events reject the null hypothesis with different node age limits for the trees following Brownian model of τ evolution, and provide support for the ortholog conjecture.

maximum speciation age), provided support for the ortholog conjecture (Fig 3), in contrast to previous results. We repeated the analysis including slightly older duplication events (duplication age ≤ 370 My, i.e. maximum 25% older than the oldest speciation). We still were able to detect higher contrasts for duplication than for speciation (Fig 3). Finally, randomization tests on these trees indicated no bias in the inference (S7A-S7D Figs), supporting the relevance of the inferences on empirical data.

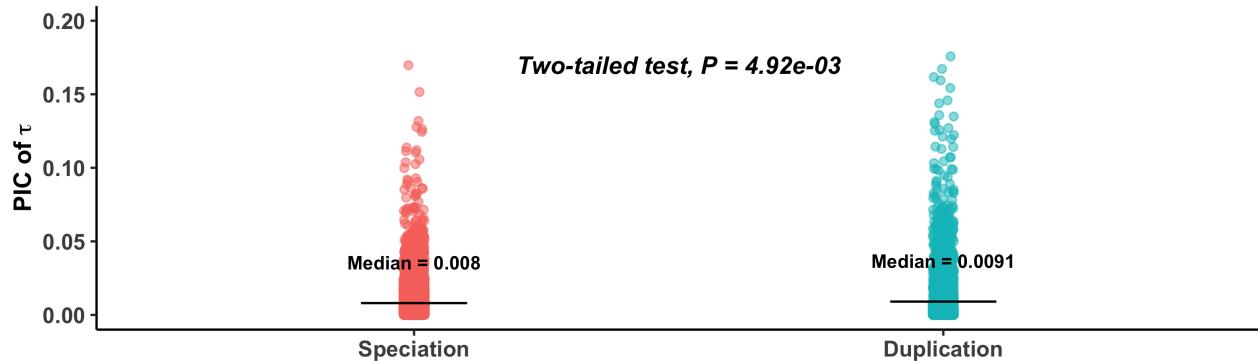
Among these 2545 trees (S6 Fig), very old duplication node contrasts were also involved to identify contrast distribution of τ as per BM, which might bias the analysis. Hence, we repeated the identification of BM trees by standard diagnostic plot tests considering only 5479 trees with at least one duplication event, after removing very old duplication nodes (duplication age > 370 My) (S6 Fig). We still were able to find support for the ortholog conjecture on these identified 2682 BM trees (S8 Fig).

To avoid calibration bias of older duplication events, we also considered a more restricted set of 806 passing more strict conditions (S6 Fig), i.e. BM trees in which there is at least one duplication, and a speciation event older than all duplications. Calibration bias for duplication events is expected to be very limited for such trees. Support for the ortholog conjecture still holds with these 806 trees (Fig 4A). Randomization tests on these 806 trees also support the biological relevance of these results (Figs 4B and 4C,S9 Fig). These observations confirm that the ortholog conjecture is supported for adequately standardized BM trees.

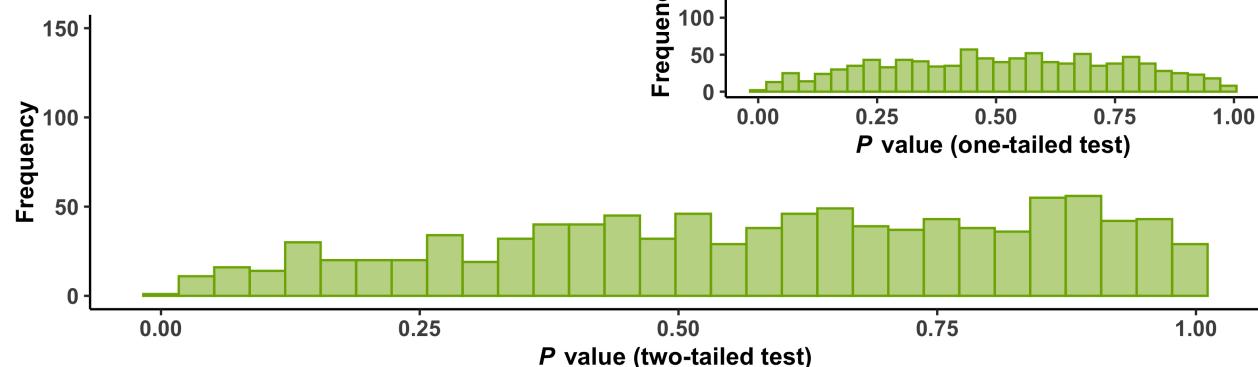
Instead of using standard diagnostic plot tests [13–15,18–20], Dunn et al. [7] considered model-fitting criteria to identify BM trees from their list of 8520 calibrated trees. They contrasted the BM model (neutral drift model with no selection) to the Ornstein-Uhlenbeck (OU) model (i.e. an extended BM model with stabilizing selection) [26,27] for this purpose. After computing the difference in Akaike Information Criterion (ΔAIC) [28] scores of the models for each tree, they identified 3151 BM trees, out of which only 8 trees purely favored the BM model ($\Delta\text{AIC}_{\text{BM}} < 2$ and $\Delta\text{AIC}_{\text{OU}} \geq 2$). The other 3143 trees favored both the BM and OU models ($\Delta\text{AIC}_{\text{BM}} < 2$ and $\Delta\text{AIC}_{\text{OU}} < 2$). A Wilcoxon two-tailed test on those 3151 trees provided similar results to the 8520 calibrated trees, i.e. higher τ evolution for orthologs than for paralogs ($\text{PIC}_{\text{speciation}} = 0.0063$, $\text{PIC}_{\text{duplication}} = 0.0059$, $P = 4.68e - 03$). This result implies that selection of BM or non-BM trees does not explain these results. A similar trend was observed for the recent duplicates (node age ≤ 296 My) (S9A Fig). The pattern of systematic increase of duplication PICs with node age (S10A Fig) plausibly implies that the

evolution of τ does not in fact follow Brownian motion for these trees. Randomization tests on those trees (S10B and S10C Figs) confirmed that contrasts were not appropriately standardized as per BM, and thus the inference drawn on empirical data was not supported. When we performed diagnostic plot tests on those 3151 trees, we found support for the ortholog conjecture for 2412 trees passing the diagnostic tests (S11 Fig). All these analyses suggest that Dunn et al. did not identify accurately BM trees, and that such misidentification leads to violating the assumptions of phylogenetic comparative methods, and to erroneous results.

A. Empirical result



B. Randomized τ



C. Randomized events

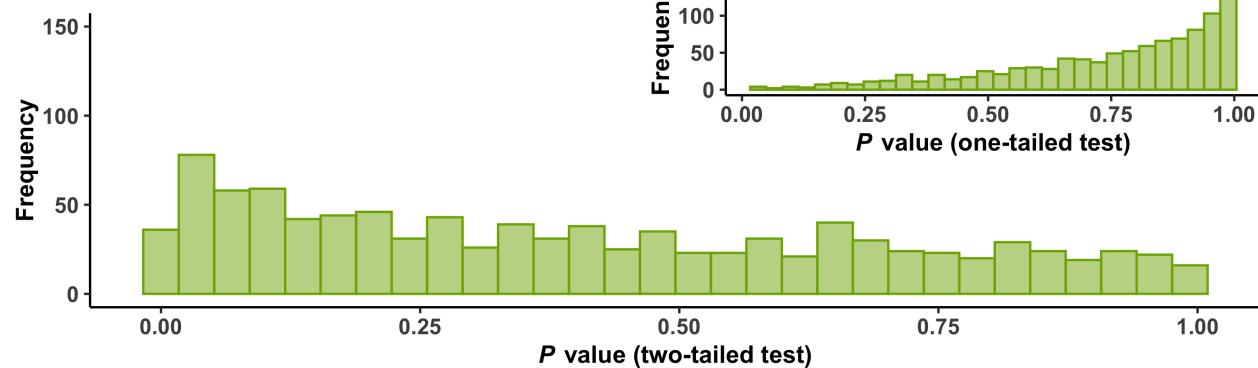


Fig 4: Results on real and randomized data for 806 BM trees with ancient speciation events. P values are from two-tailed Wilcoxon tests. (A) The jitter plot on empirical gene tree data supports the ortholog conjecture. Main plots in (B) and (C) show the distributions of two-tailed test P values estimated for the whole set of 806 BM trees by permuting τ (B), and by permuting the internal node events (C) ('Speciation', 'Duplication', and 'NA') for 1000 independent runs respectively. Inset plots in (B) and (C) show P value distributions plots of one-tailed test to check if there was an excess of duplication contrasts over speciation contrasts by randomizing τ , and events respectively. The proportions of speciation (median = 0.62), and NA events (median = 0.2) are much higher than that of duplication events (median = 0.14) in these trees (S9 Fig). Hence, higher duplication contrasts are more likely to be replaced by speciation events by permutations of events. Due to this reason, we found significant shifts towards left in the main plot of C, and towards P value 1 in the inset plot of C to depict larger contrasts for speciation events. Comparisons of plot A to plot B, and plot A to plot C show that randomization results in both the cases actually differ from the empirical result.

Discussion

We agree with Dunn et al. [7] that evolutionary comparisons should be done considering a phylogenetic framework when possible. However, this does not imply that phylogenetic independent contrasts can be applied easily to phylogenomics data and questions. To get a clear picture, we limited our study to the same phylogenetic gene trees used by Dunn et al. [7]. Our reanalysis identified problems generated during the time calibration of duplication nodes of pruned trees for which no external time references are available, and with non respect of the hypothesis of Brownian motion. The strongest bias was for duplication nodes preceding the oldest speciation nodes. This, in turn, introduced several biases in their analyses, and influenced their results. When we identified and controlled for such biases, PIC results changed to support the ortholog conjecture, consistent with our previous pairwise approach [6] on the same τ data.

To our knowledge, no one before Dunn et al. [7] applied a phylogenetic approach in comparative biology to study the effect of gene duplication on functional evolution, despite an early call to do so [29]. In line with previous studies [13,30], we explored whether the application of a phylogenetic method might inflate errors (e.g. rejection of the null hypothesis in null condition, Figs 1A and 1B) if applied without thorough testing for the fundamental assumptions of the method. Assumptions of proper branch length information and of Brownian model of trait evolution are related, so that modifications of branch lengths can change the evolutionary model [30]. We find that branch length calibration is mostly inaccurate for old duplication events (duplication events prior to the oldest speciation event in the data) (S2 Fig). Use of such node contrasts causes a strong rise in expected variance for duplication events compared to the speciation events (S4 Fig). This may bring about lack of statistical power to detect the signal of ortholog conjecture, when in reality the signal is present, and even bias towards a pseudo-signal. As a remedial measure, we limited our analysis to the trees for which Brownian motion of trait evolution has been identified with the aid of standard diagnostic plot tests [13–15,18–20]. To remove issues with branch length inaccuracies for older duplication events, we used limitation of node ages for duplication events. Using all such measures to control for biases, we found support for the ortholog conjecture (Figs 3 and 4A). The reliability of our inference was validated by two different randomization tests, which confirmed the fact that our result was not due to biases in the data or analysis (S7A-S7D Figs, Figs 4B and 4C).

Being aware of the fact that branch lengths are fundamental to phylogenetic comparative methods, Dunn et al. [7] added random noise in the speciation calibration time point, and extended terminal branch lengths to test sensitivity of their observations. While extending terminal branch can reduce error rate due to the presence of negative branch lengths in 54 trees, none of the modifications appear to avoid biases when there are so vast differences in branch lengths [31–37]. These can also lead to violating the Brownian model of trait evolution [30]. Diagnostic plot tests appear necessary to assess adequate standardization of contrasts as per BM in such cases [13–15,18–20].

The importance of proper BM tree selection should not be underestimated as it can impact the contrast analysis. There exist no fixed protocols to select BM trees. We used standard diagnostic plots recommended in earlier studies for PIC [13–15,18–20,38–40] to rule out significant departure from BM for each tree individually at 95% confidence level. In contrast, Dunn et al. [7] used both BM and OU as null models for each tree, leading them to identify “BM trees” for which the two models were equally good fits. These do not appear to be proper BM trees according to several observations, since duplication contrasts increased with node age (S10A Fig). Moreover, the inferences based on such trees with empirical data were not different from that based on randomized trees (S10B and S10C Figs). When we applied diagnostic tests on those 3151 “BM trees”, 734 trees failed the diagnostic test, and rest of the trees provided support for the ortholog conjecture (S11 Fig).

Empirical support for the ortholog conjecture has been mixed, with most studies supporting, and a few failing to do so [3–6,8]), our results provided support for the ortholog conjecture using large-scale genome wide tissue specificity data in a phylogenetic framework after controlling for bias. Due to lack of detailed functional information, many studies are still limited to gene expression data as a proxy of function. Recently, using functional replaceability assay, experimental studies [41,42] have shown that orthologous genes can be swapped between essential yeast genes and human, although this is rarely the case for all the members of expanded human gene families [42], validating one prediction of the ortholog conjecture.

These analyses are mainly based on gene trees dominated by small scale duplication events. In these trees, the age of the same duplication clade is never fixed. Use of time calibrated whole genome duplication trees could be beneficial in this regard. Since the approximate time period of whole genome duplications are known, we can test the hypothesis more conveniently by avoiding use of any node age limits on whole genome duplication events for those trees.

Methods

Resource details for reanalyses

Our reanalyses are based on 8520 annotated (with the events: speciation, duplication and NA), pruned, and time calibrated ENSEMBL Compara v.75 [43] gene trees, having at least 4 tips with non null trait data. They were obtained from Dunn *et al* [7]. Like them [7], we used a subset of precomputed τ data (as trait) of 8 vertebrate species from the study of Kryuchkova-Mostacci and Robinson-Rechavi [6]. Kryuchkova-Mostacci and Robinson-Rechavi [6] computed τ and mean gene expression levels by following the method of Yanai et al. [44]. The subset of data was based on the RNA-seq data provided by Brawand et al. [45] for 6 tissues. We used the same random seed number as in [7] to reproduce the simulation results for reanalysis. All reproduced data of Dunn et al. were stored in the “manuscript_dunn.RData” file (<https://doi.org/10.5281/zenodo.3354285>). We used the results stored in the ‘data’ or ‘phylo’ slot of the trees for further analyses. To differentiate our own function from theirs [7], we remodified the original function script of Dunn et al. from “functions.R” to “functions_Dunn.R”. Some of the analyses were time consuming, so we made a separate script “Premanuscript_run_TM.R” to run before knitting the markdown file. We stored the outputs in ???Data_TMRR.rda??? file (<https://doi.org/10.5281/zenodo.3354285>) and loaded it during our analyses. All the details of different functions were provided inside the script. We supply all the previously stored data (to reduce computation time during reproduction of result) and function files including our own (“function_TM_new.R”) with this manuscript. All scripts are available on GitHub: https://github.com/tbegum/Testing_the_ortholog_conjecture.

We used phylosig function() of “phytools” package [46] to identify trees with phylogenetic signal ($P < 0.05$) using Blomberg’s K [22,23,46].

Analyses and plotting were performed in R version 3.5.1 [40] using treeio [47], geiger [48], stringr [49], digest[50], dplyr [51], tidyverse [52], ggrepel[53], gtools [54], ggplot2 [55], cowplot [56], easyGgplot2 [57], gridExtra [58], and png [59] libraries.

Randomization test of τ values

For each tree, we used τ data (column name “Tau” in each tree ‘data’ object) across the tips to carry out our randomization test. To randomize we permuted the actual τ data without altering internal node events. The pic() function of the “ape” package [60] was used to compute PIC of nodes for each tree using permuted τ of tips. For each run, we compared the contrasts of speciation and duplication events of the whole set of randomized trees to estimate difference in event contrasts based on Wilcoxon signed rank test. For 100 or 1000 runs, we repeated the above process 100 or 1000 times to obtain a distribution plot of 100 or 1000 independent P values.

Randomization test of node events

Some of the speciation nodes had daughters with same clade names in the gene trees we used for our study. Dunn et al. changed such node events to “NA” to avoid problems during time calibration of the trees. Such annotated node event information (“Speciation”, “Duplication”, “NA”) for each tree was available as “Event” in the tree ‘data’ slot. To randomize, we permuted the internal node events (added as column name “event_new” in the ‘data’ slot) by maintaining the actual proportion of events for each tree. Then, we used the PIC of actual τ at tips to estimate contrasts difference between newly assigned speciation and duplication node events by Wilcoxon rank tests. For 100 or 1000 independent runs, we repeated the same procedure to obtain 100 or 1000 independent P values.

Checking for Brownian model of trait evolution

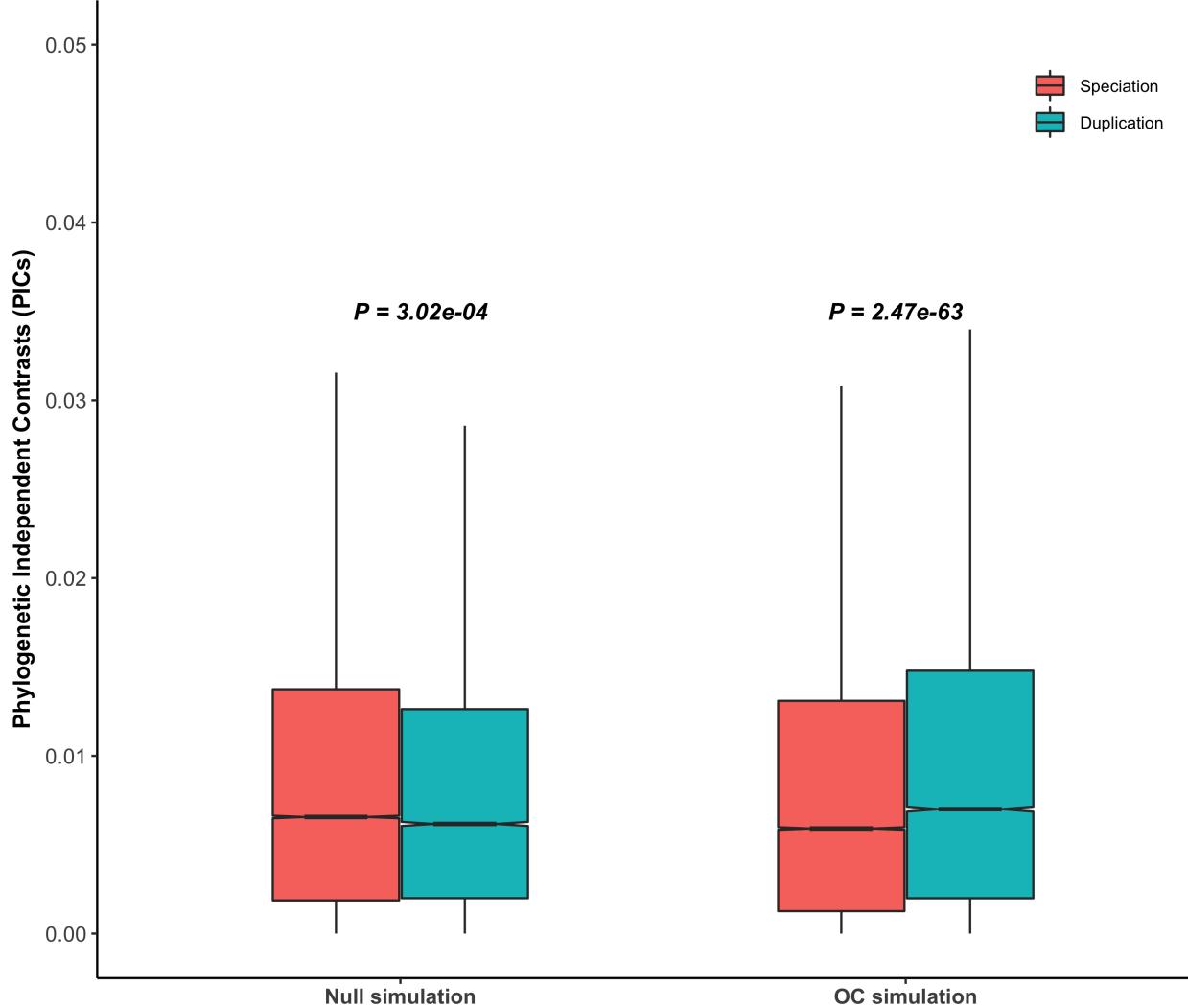
There exists no established protocol to identify the BM (Brownian model) trees. Dunn et al. used model-fitting with $\Delta AIC_{BM} = AIC_{BM} - \min(AIC_{BM}, AIC_{OU})$ and $\Delta AIC_{OU} = AIC_{OU} - \min(AIC_{BM}, AIC_{OU})$. They applied $\Delta AIC_{BM} < 2$ as the relative support for the BM model. We used several diagnostic tests to test BM behavior, as recommended in several studies [13–15,18,19,30]. Among diagnostic tests, the most usual method for contrasts standardization is to check a correlation between the absolute values of standardized contrasts and their expected standard deviations [13,18,20]. Under Brownian motion, there should be no correlation. This

test and the correlation between the absolute values of standardized contrasts and the logarithm of their node age are model diagnostic plot tests in the *caper* (“Comparative Analyses of Phylogenetics and Evolution in R”) package [18,38–40]. We used both of them in our study by using the “crunch” algorithm of the “caper” package in R, which implements the methods originally provided in CAIC [18,38–40]. Correlation of node heights with absolute values of contrasts has also been reported to be a reliable indicator of deviation from the Brownian model [19]. Hence, we computed node height for each node in a tree using the *ape* package [60]. We also used the correlations of node height and node depth to the absolute value of nodes contrasts to rule out significant trend in any of the 4 tests. We used $P < 0.05$ to assess a significant correlation for the diagnostic plot tests. A significant trend (positive or negative) reliably indicates a deviation from the BM of trait evolution for that tree [13–15,18–20], and we removed those trees from our analysis. Contrast calculation on negative branch lengths is not desirable, so we removed trees with negative branch lengths before applying the *crunch()* function. Trees passing all 4 diagnostic tests were considered as BM trees for our study.

Acknowledgements

We sincerely acknowledge Martha Liliano Serranno Serranno for initial help in understanding the phylogenetic independent contrast method. We thank Nicolas Salamin, Julien Wollbrett, Jialin Liu, Sebastien Moretti, Sara Fonseca Costa, Kamil Jaron and all the members of the Robinson-Rechavi group for their help and useful discussions. We also acknowledge Cassey Dunn for his initial help in reproducing their results. Parts of the computations were performed at the Vital-IT (<http://www.vital-it.ch>) Center for high-performance computing of the SIB Swiss Institute of Bioinformatics.

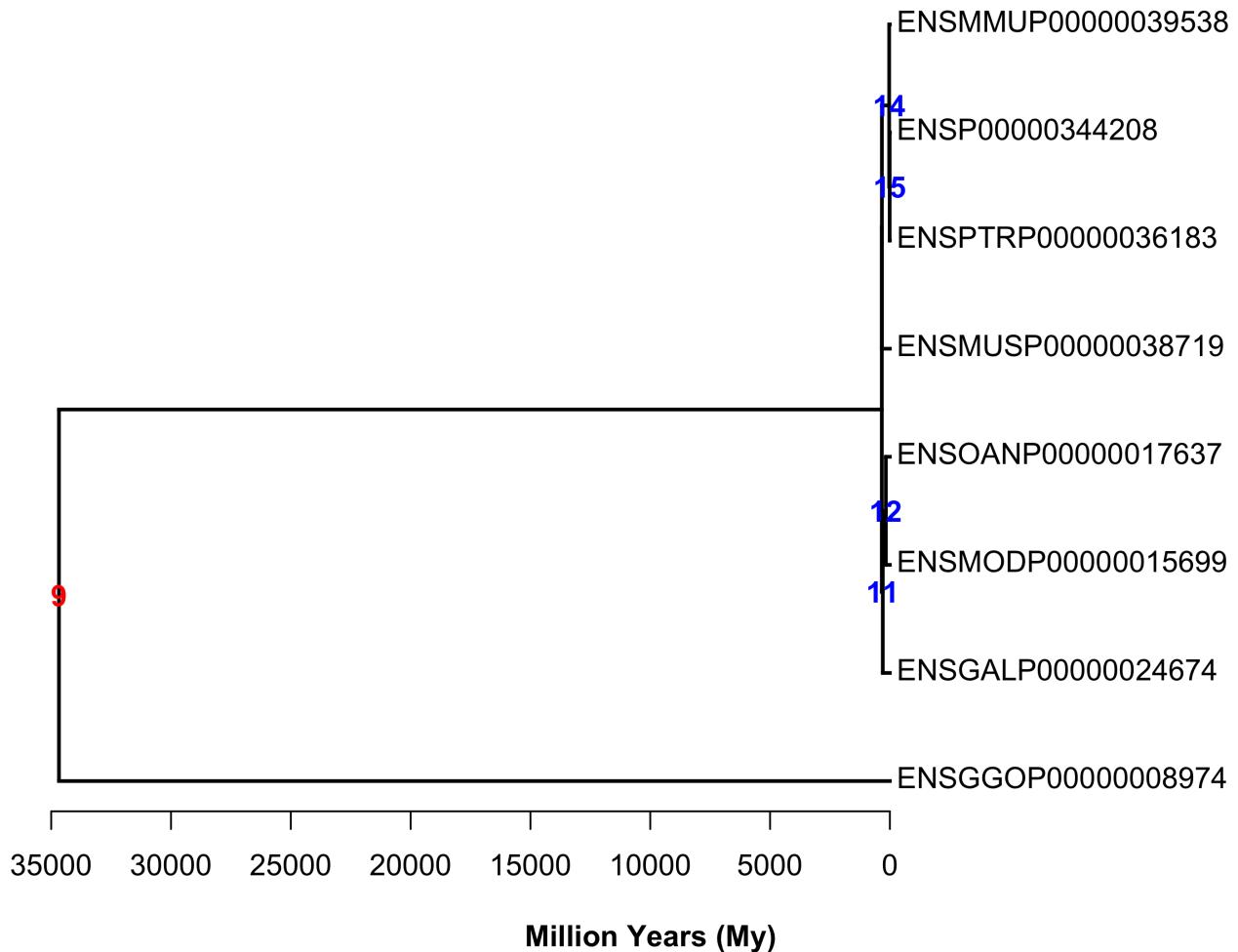
Plot on 8520 calibrated trees



S1 Fig: Repeating simulations on all calibrated trees with different random seed number. P values are from Wilcoxon two-tailed tests. Simulations with different seed number did not change the trend of results as observed in Fig 1A.

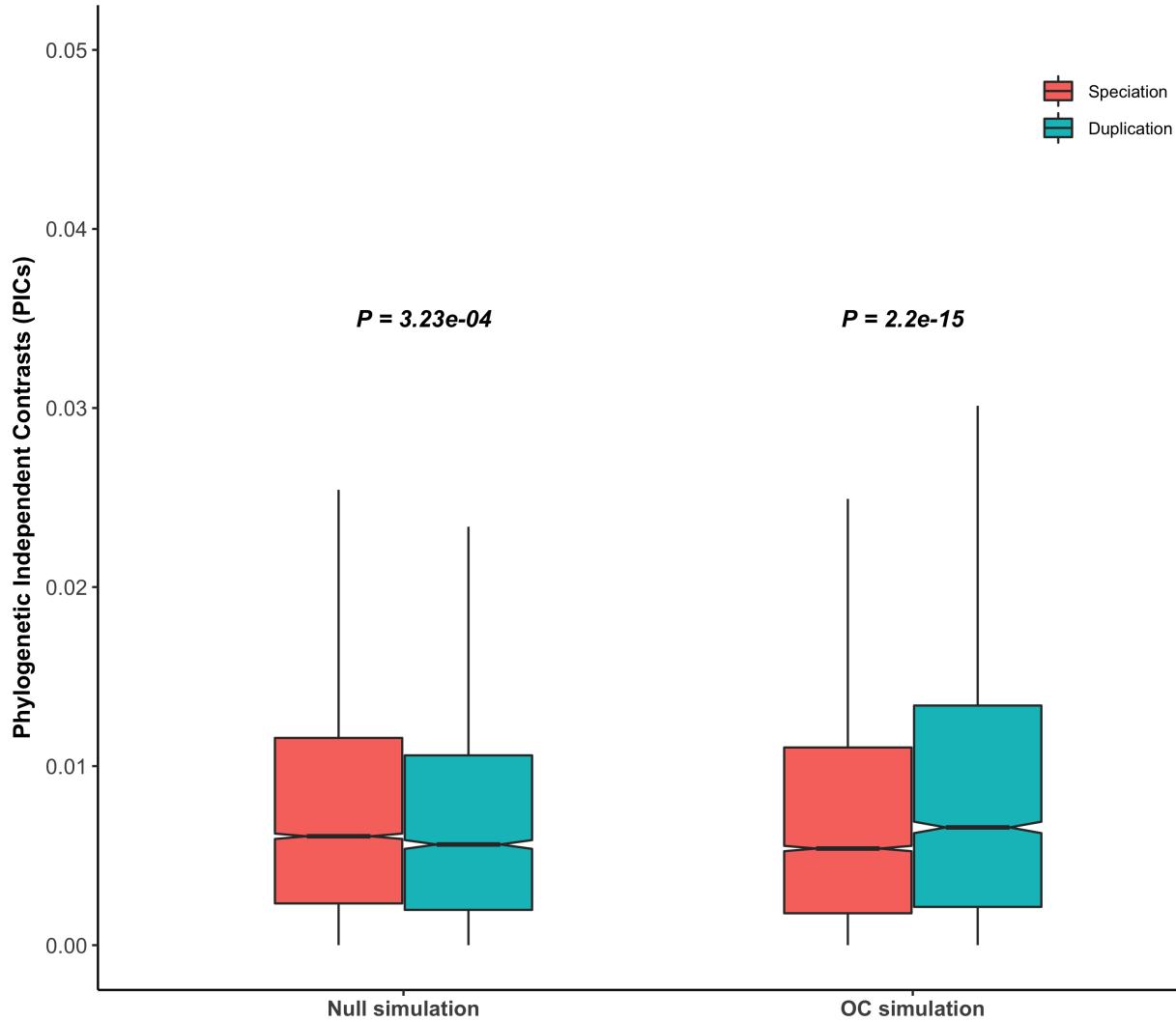
Supporting Information

An example of time calibrated trees for two different events

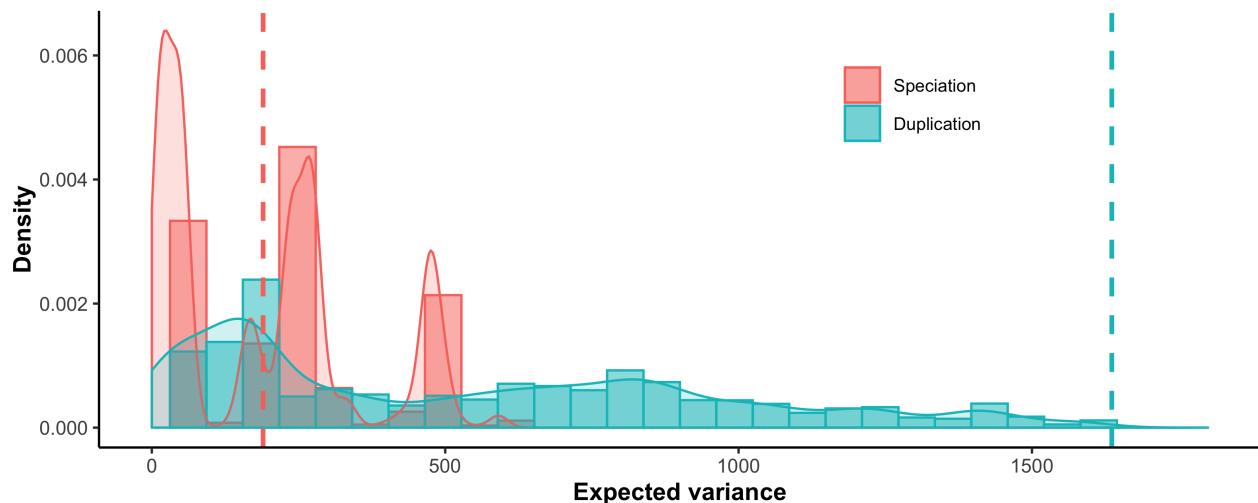
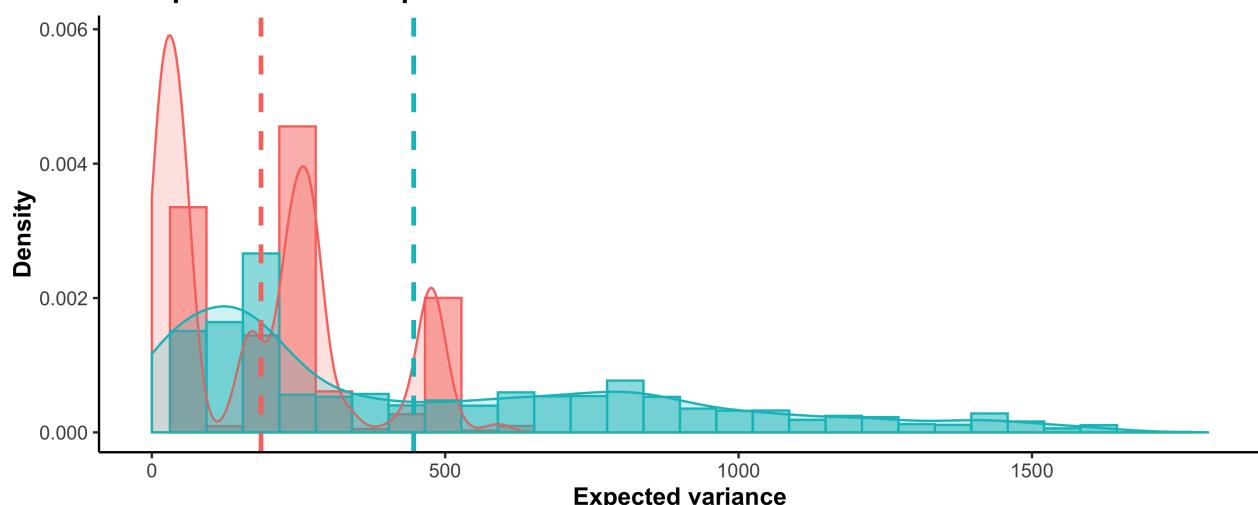


S2 Fig: An example of a calibrated gene trees used by Dunn et al. in their study. The blue nodes represent the speciation nodes while the red nodes are duplication nodes. For nodes with no number the event is assigned as NA by the authors [7]. Since, the duplication time is unknown, the tree is calibrated based on speciation node ages. Notice that this gives an unreasonably large estimated date for the ancient duplication, resulting in higher expected variance and abnormally low contrast for the corresponding node event. This example is to demonstrate that it is often risky to use such nodes to infer the result of phylogenetic method that intends to study the effect of gene duplication.

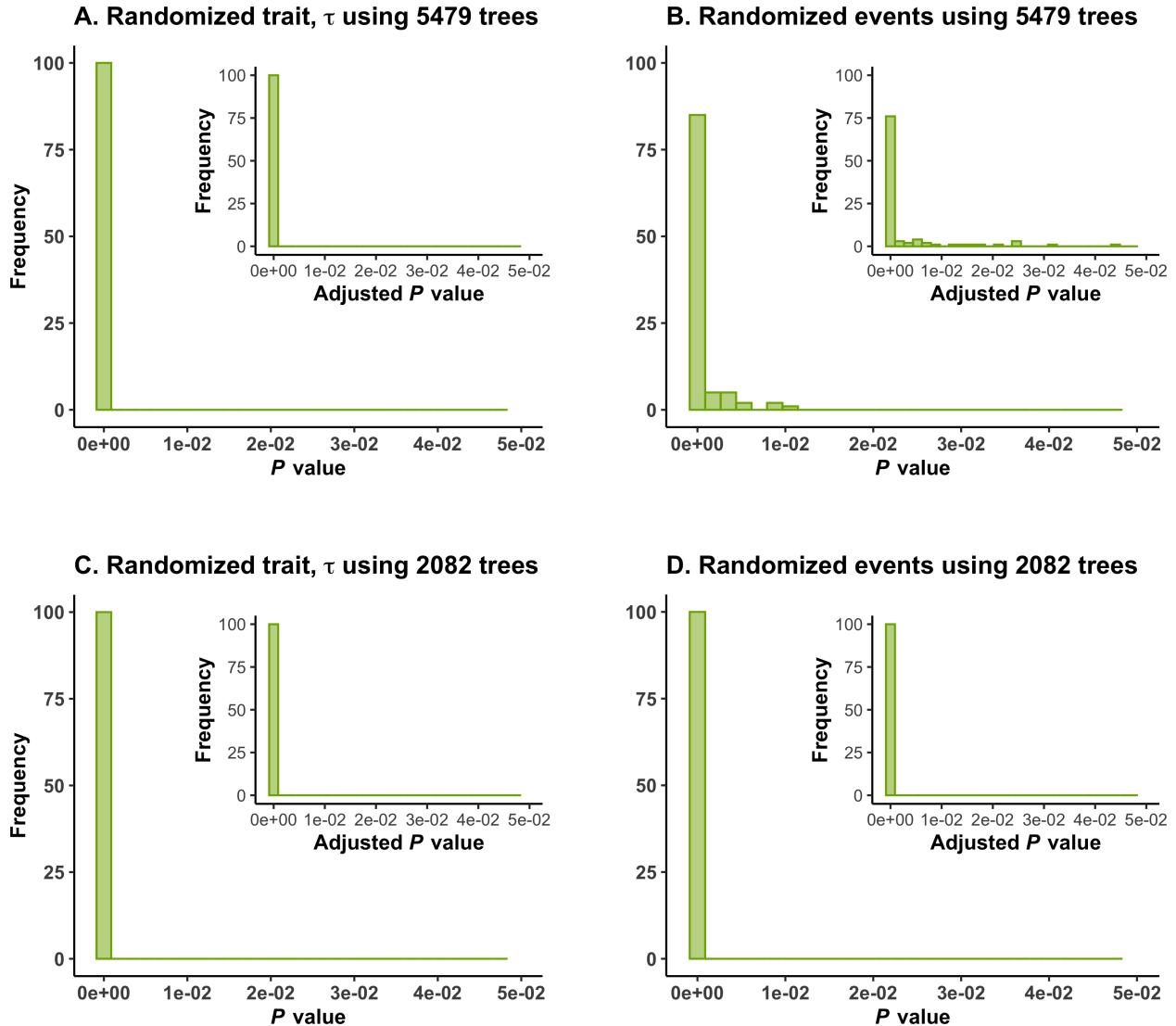
Plot on trees with strong phylogenetic signals



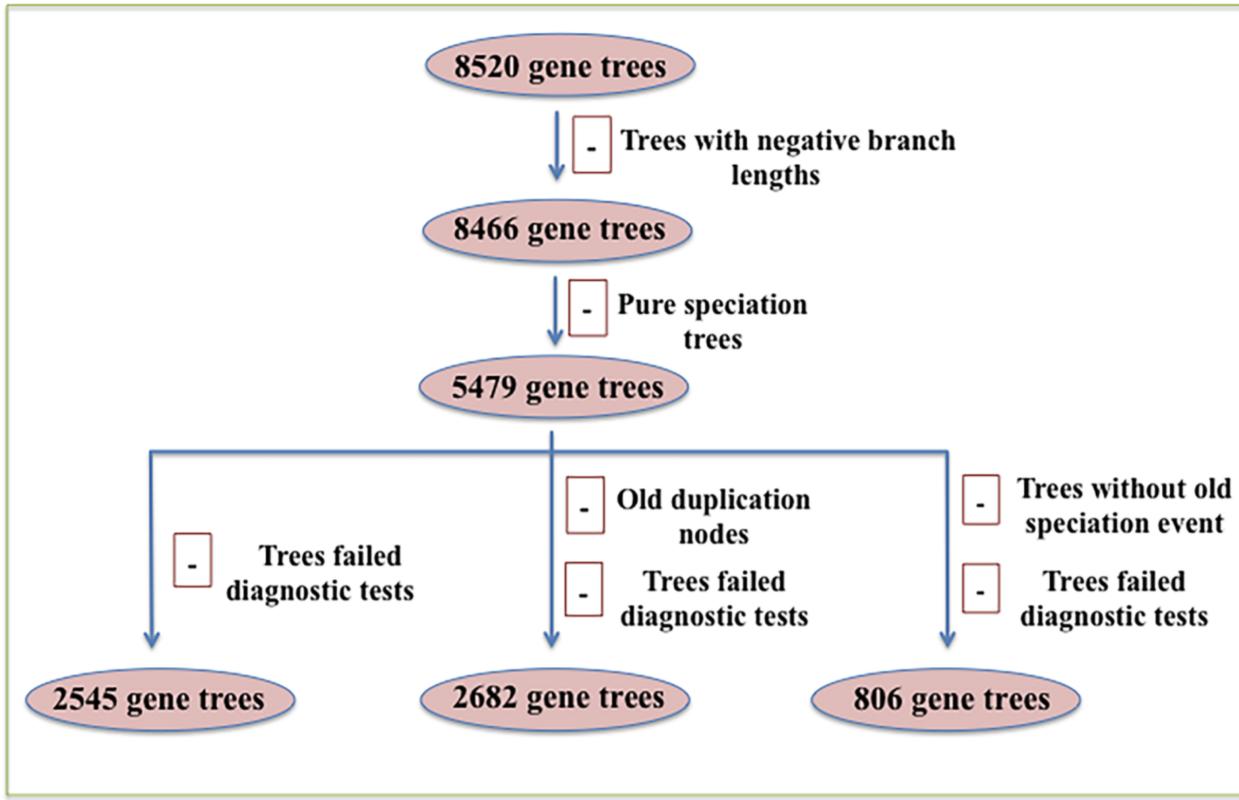
S3 Fig: Simulation analyses on 1135 trees with strong phylogenetic signals. P values are from Wilcoxon two-tailed tests. Dunn et al. used a cutoff of $K > 0.551$ to identify trees with strong phylogenetic signals. However, trees with higher K statistic can reflect trees with no phylogenetic signal when corresponding P values are non-significant. Considering both K statistic and P value, we found similar trends as was observed with 2082 trees.

A. Expected variance plot of 8520 trees**B. Expected variance plot of 2082 trees**

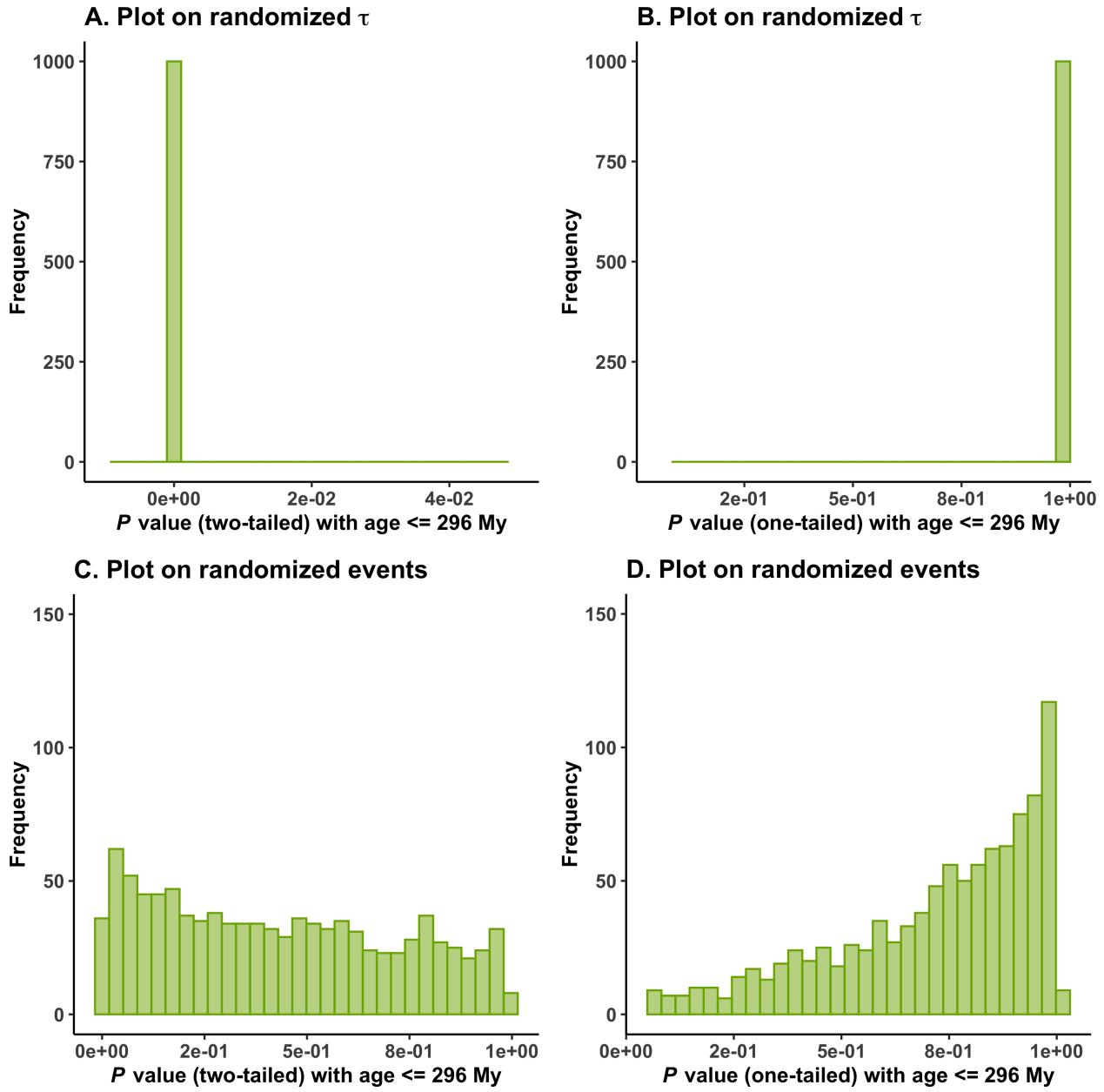
S4 Fig: Re-analyses of expected variances of calibrated trees considered by Dunn et al. The expected variance plots of (A) all calibrated trees, and (B) trees with strong phylogenetic signal. The dotted line represents the mean expected variance of the events. These plots suggest that the use of duplication nodes preceding ancient speciation nodes for calibration is often erroneous and should be avoided.



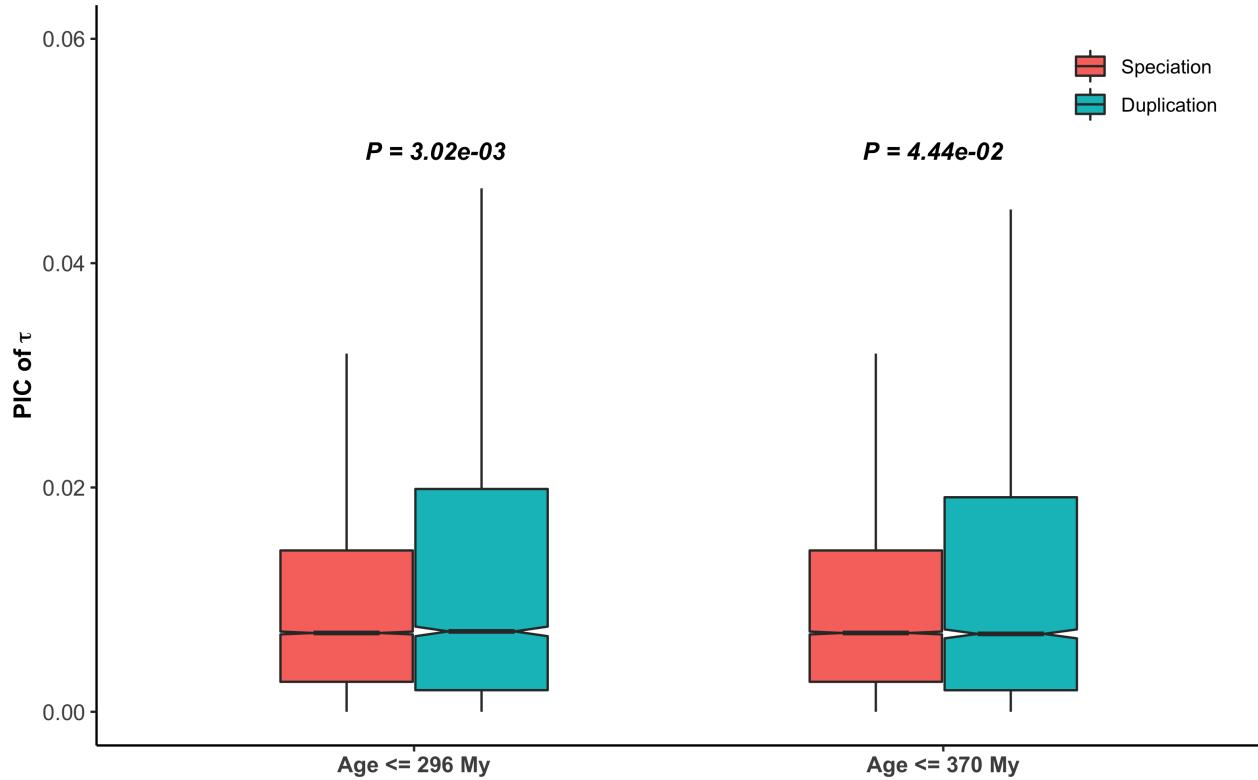
S5 Fig: P value distribution plots after 100 independent runs on each set of trees. Wilcoxon two-tailed test with 95% confidence interval was used to compare the speciation and duplication contrasts after randomization tests. (A) and (B) applied to trees with at least one speciation and one duplication event. (C) and (D) applied to trees with strong phylogenetic signal. (A) and (C) randomization of trait (τ) over the trees. (B) and (D) randomization of internal node events. The inset plots show P values adjusted with Benjamini-Hochberg [61,62]. Supporting our observations of Figs 2A and 2C, all the plots confirm that the empirical result of Dunn et al. [7] is not different from randomized test results.



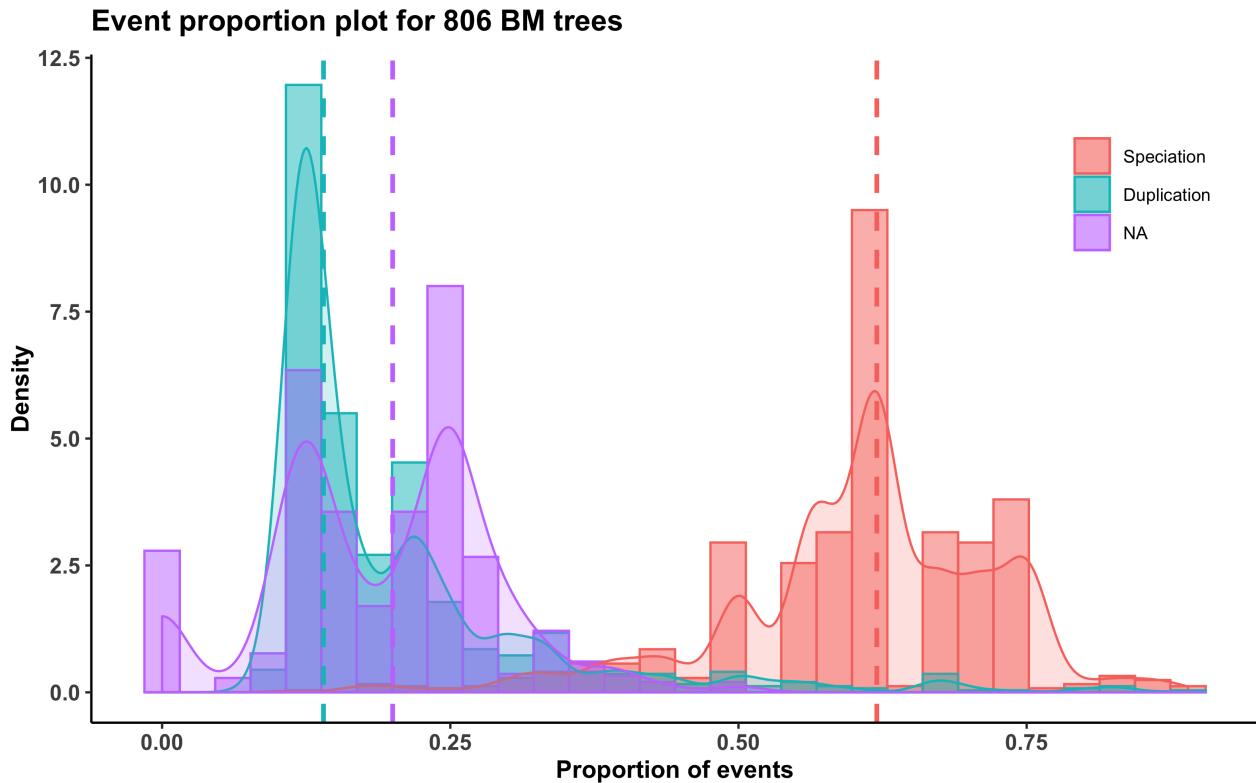
S6 Fig: Pipeline to generate different subsets of contrasts standardized BM trees of τ . The symbol ‘-’ indicates removal of trees with the corresponding mentioned criteria. Diagnostic plot tests (discussed in details in Methods) refer to the tests recommended in prior studies [13–15,18–20] to identify trees where trait (in this case τ) evolution follows BM. Old duplication nodes refers to duplication events older than 370 My. Pure speciation trees indicate gene trees with no duplication event. Removal of trees without old speciation means that we keep only trees where all duplication events are more recent than the oldest speciation event, i.e. trees with age \leq 296 My.



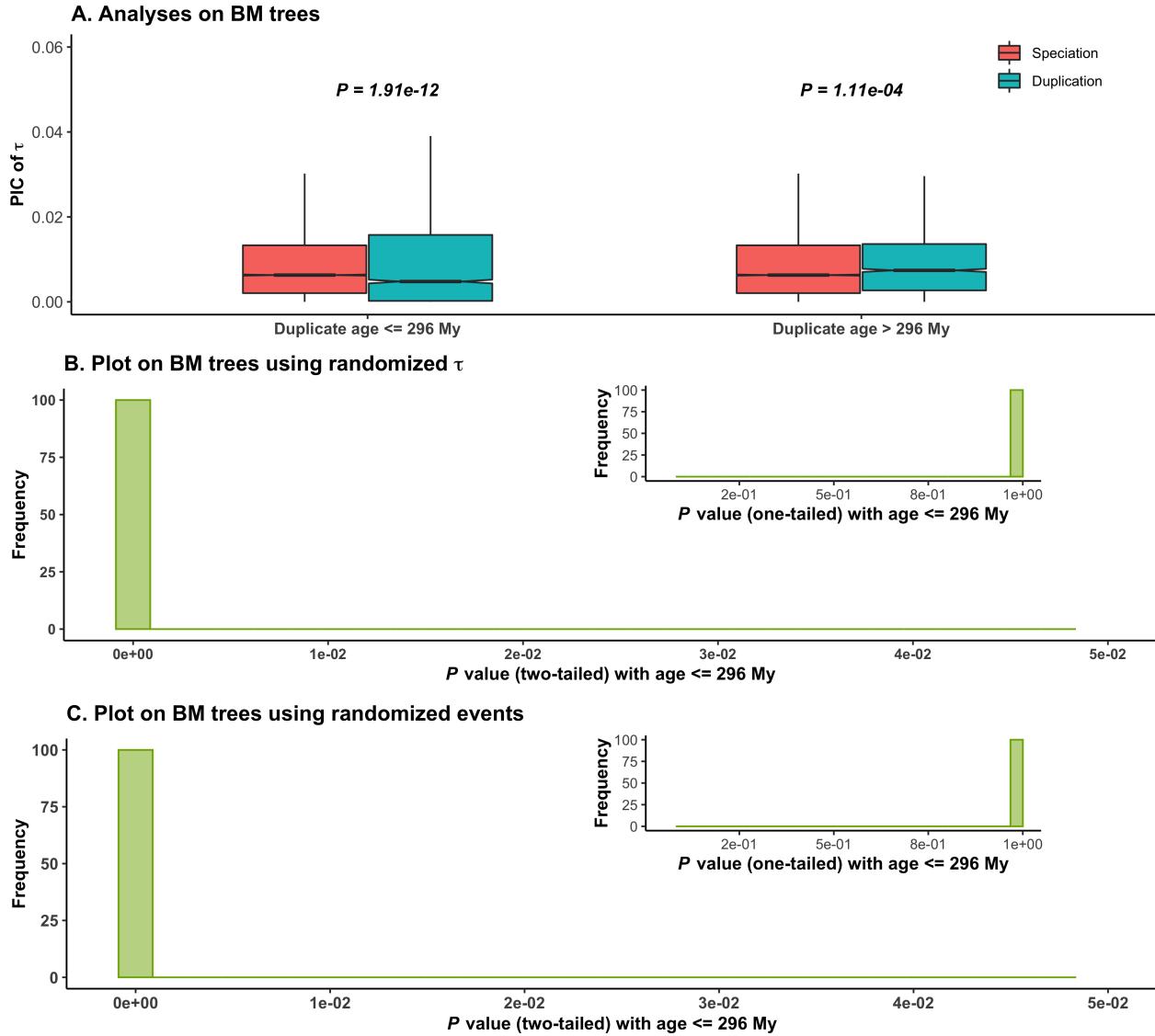
S7 Fig: Randomization test results on contrasts standardized 2545 BM trees for τ . Wilcoxon tests were used to generate P value distribution plots by 1000 independent runs with a node age limit of 296 My. (A) For randomized τ , Wilcoxon two-tailed tests unexpectedly showed all the values were significant. (B) P value of 1 in all cases of Wilcoxon one-tailed test (as performed in Dunn et al.) on randomized τ . This may be due to the fact that there is an excess proportions of speciation and NA events compared to the duplication events in these trees. By permuting τ , higher τ values of duplication events are more likely to be assigned to the speciation and NA events. Hence, this produced higher speciation contrasts, although we controlled the bias in expected variance of older duplication events by node age limits. (C) and (D) Wilcoxon two-tailed and one-tailed tests were performed by randomization of events. After randomization, the smaller proportions of duplication events were more likely to be replaced by speciation and NA events, showing greater shifts towards excess speciation contrasts. These plots show that randomization results are different from the empirical results of Fig 3, as expected.



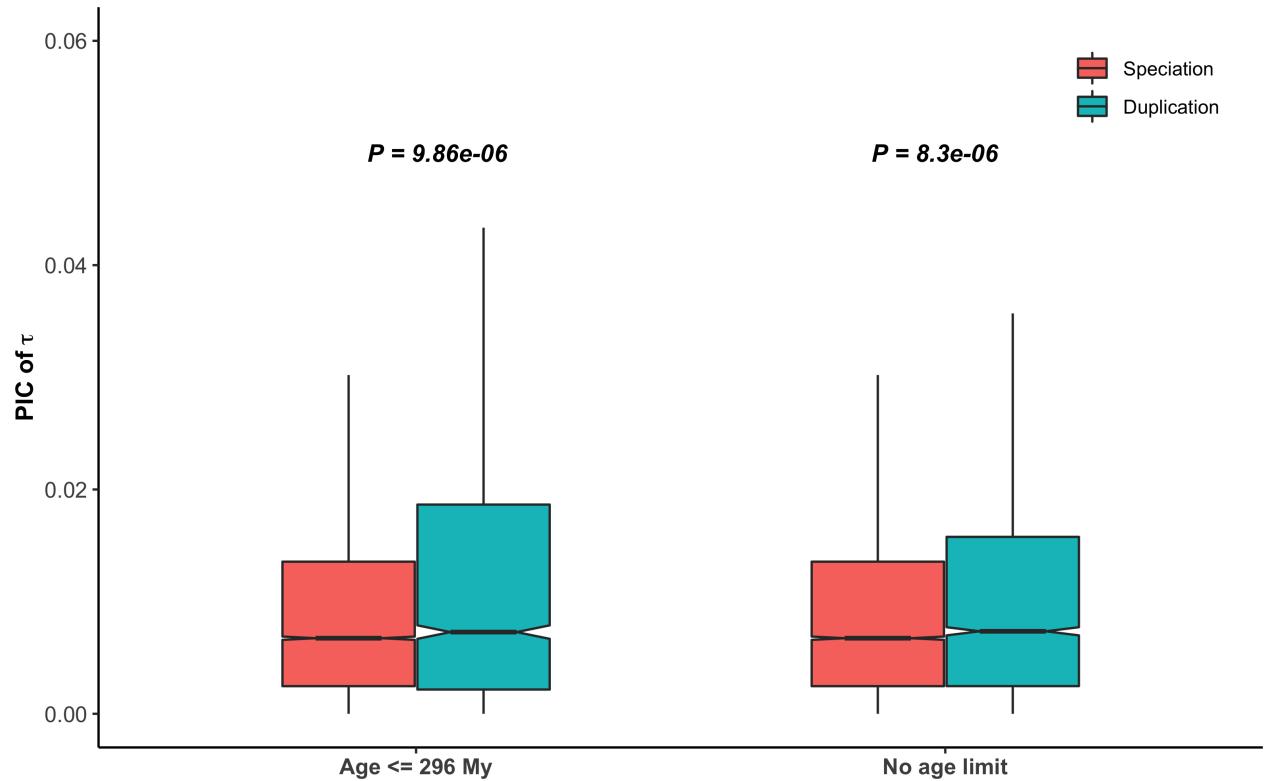
S8 Fig: Result on 2682 BM trees using maximum age of oldest duplication event as 370 My. Wilcoxon two-tailed test was used to estimate the significance level at 95% confidence interval. With different node age limits (as used in Fig 3), we found support for the ortholog conjecture. The difference between S8 Fig and Fig 3 is that in Fig 3 we used the PICs of all the old duplication nodes to identify BM trees, and then we used different node age limits to test the ortholog conjecture, whereas in S8 Fig we removed the old duplication nodes (duplication age > 370 My) before identifying BM trees, and tested for the ortholog conjecture with different node age limits.



S9 Fig: Distribution of proportions of events in 806 BM trees. These trees have a median of 8 internal nodes. Most of the trees have speciation nodes whose descendants have the same clade name. Hence, Dunn et al. [7] changed such nodes from speciation events to NA to avoid problem with tree calibration. From the figure, it is clear that the proportion of speciation events is much higher than that of the duplication events. For these trees, the proportions of NA events (median: 0.2) are even higher than the proportion of duplication events (median: 0.14). The dotted line represents the median proportions of the events.



S10 Fig: Re-analyses of 3151 trees called as following a BM model by Dunn et al. (A) Wilcoxon two-tailed test was used to obtain the significance level using empirical data. The box plot shows that there is no support for the ortholog conjecture for recent duplicates ($age \leq 296$ My), but that there is support for the older duplicates. (B) and (C) P values estimated using randomized τ (B), and using randomized events (C), with a node age limit of 296 My to test if there was any difference in contrasts between speciation and duplication events, and plotted based on 100 independent P values. The main plots in (B) and (C) demonstrate that speciation and duplication contrasts differ significantly in each randomization test. The inset plots of (B) and (C) with Wilcoxon one-tailed test to check if duplication contrasts was higher than the speciation contrasts, confirm the opposite trend, as was observed with empirical data with node age ≤ 296 My. This suggests that contrasts are not adequately standardized for these trees as per BM, and that the inference based on these trees can be misleading.



S11 Fig: The ortholog conjecture test on Dunn et al. identified BM trees, after diagnostic tests. Among 3151 trees, 2412 trees passed the diagnostic tests of BM. Wilcoxon two-tailed test was used to estimate the significance level at 95% confidence interval. We did not exclude pure speciation trees before the contrasts analyses. With node age limit (age ≤ 296 My) as well as without, we found support for the ortholog conjecture.

Reference

1. Studer RA, Robinson-Rechavi M (2009) How confident can we be that orthologs are similar, but paralogs differ? *Trends Genet* 25: 210–216. Available: <https://www.ncbi.nlm.nih.gov/pubmed/19368988>.
2. 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. Available: <https://www.ncbi.nlm.nih.gov/pubmed/21695233>.
3. 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. Available: <https://www.ncbi.nlm.nih.gov/pubmed/22615551>.
4. 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. Available: <https://www.ncbi.nlm.nih.gov/pubmed/23209392>.
5. Rogozin IB, Managadze D, Shabalina SA, Koonin EV (2014) Gene family level comparative analysis of gene expression in mammals validates the ortholog conjecture. *Genome Biol Evol* 6: 754–762. Available: <https://www.ncbi.nlm.nih.gov/pubmed/24610837>.
6. 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. Available: <https://www.ncbi.nlm.nih.gov/pubmed/28030541>.
7. Dunn CW, Zapata F, Munro C, Siebert S, Hejnol A (2018) Pairwise comparisons across species are problematic when analyzing functional genomic data. *Proc Natl Acad Sci U S A* 115: E409–E417. Available: <https://www.ncbi.nlm.nih.gov/pubmed/29301966>.
8. Gabaldón T, Koonin EV (2013) Functional and evolutionary implications of gene orthology. *Nat Rev Genet* 14: 360–366. Available: <https://www.ncbi.nlm.nih.gov/pubmed/23552219>.
9. Koonin EV (2005) Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* 39: 309–338. Available: <https://www.ncbi.nlm.nih.gov/pubmed/16285863>.
10. Sonnhammer EL, Gabaldón T, Sousa da Silva AW, Martin M, Robinson-Rechavi M, et al. (2014) Big data and other challenges in the quest for orthologs. *Bioinformatics* 30: 2993–2998. Available: <https://www.ncbi.nlm.nih.gov/pubmed/25064571>.
11. Felsenstein J (1985) Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39: 783–791. Available: <https://www.ncbi.nlm.nih.gov/pubmed/28561359>.
12. Felsenstein J (1985) Phylogenies and the comparative method. *The American Naturalist* 125: 1–15. Available: <https://www.jstor.org/stable/2461605>.
13. Díaz-Uriarte R, Garland T (1998) Effects of branch length errors on the performance of phylogenetically independent contrasts. *Syst Biol* 47: 654–672. Available: <https://www.ncbi.nlm.nih.gov/pubmed/12066309>.
14. Garland T (1992) Rate tests for phenotypic evolution using phylogenetically independent contrasts. *Am Nat* 140: 509–519. Available: <https://www.ncbi.nlm.nih.gov/pubmed/19426053>.
15. Garland T, Bennett AF, Rezende EL (2005) Phylogenetic approaches in comparative physiology. *J Exp Biol* 208: 3015–3035. Available: <https://www.ncbi.nlm.nih.gov/pubmed/16081601>.
16. Grafen A (1989) The phylogenetic regression. *Philos Trans R Soc Lond B Biol Sci* 326: 119–157. Available: <https://www.ncbi.nlm.nih.gov/pubmed/2575770>.
17. Martins EP, Garland T (1991) Phylogenetic analyses of the correlated evolution of continuous characters: A simulation study. *Evolution* 45: 534–557. Available: <https://www.ncbi.nlm.nih.gov/pubmed/28568838>.
18. Cooper N, Thomas GH, FitzJohn RG (2016) Shedding light on the ‘dark side’ of phylogenetic comparative methods. *Methods Ecol Evol* 7: 693–699. Available: <https://www.ncbi.nlm.nih.gov/pubmed/27499839>.

19. Freckleton RP, Harvey PH (2006) Detecting non-brownian trait evolution in adaptive radiations. PLoS Biol 4: e373. Available: <https://www.ncbi.nlm.nih.gov/pubmed/17090217>.
20. Garland TJ, Harvey P, Ives A (1992) Procedure for the analysis of comparative data using phylogenetically independent contrasts. Systematic Biology 41: 18–32.
21. Molina-Venegas R, Rodríguez M (2017) Revisiting phylogenetic signal; strong or negligible impacts of polytomies and branch length information? BMC Evol Biol 17: 53. Available: <https://www.ncbi.nlm.nih.gov/pubmed/28201989>.
22. Münkemüller T, Lavergne S, Bzeznik B, Dray S, Jombart T, et al. (2012) How to measure and test phylogenetic signal. Methods in Ecology and Evolution 3: 743–756. doi:doi: 10.1111/j.2041-210X.2012.00196.x.
23. Blomberg SP, Garland T, Ives AR (2003) Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. Evolution 57: 717–745. Available: <https://www.ncbi.nlm.nih.gov/pubmed/12778543>.
24. Pagel M (1999) Inferring the historical patterns of biological evolution. Nature 401: 877–884.
25. Freckleton R, Harvey P, Pagel M (2002) Phylogenetic analysis and comparative data: A test and review of evidence. American Naturalist 160: 712–726.
26. Hansen TF (1997) Stabilizing selection and the comparative analysis of adaptation. Evolution 51: 1341–1351. Available: <https://www.ncbi.nlm.nih.gov/pubmed/28568616>.
27. Chen J, Swofford R, Johnson J, Cummings BB, Rogel N, et al. (2019) A quantitative framework for characterizing the evolutionary history of mammalian gene expression. Genome Res 29: 53–63. Available: <https://www.ncbi.nlm.nih.gov/pubmed/30552105>.
28. Akaike H (1974) New look at statistical-model identification. Automatic Control, IEEE Transactions on 19: 716–723. doi:10.1109/tac.1974.1100705.
29. Eisen JA, Wu M (2002) Phylogenetic analysis and gene functional predictions: Phylogenomics in action. Theor Popul Biol 61: 481–487. Available: <https://www.ncbi.nlm.nih.gov/pubmed/12167367>.
30. Diaz-Uriarte R, Garland T (1996) Testing hypotheses of correlated evolution using phylogenetically independent contrasts: Sensitivity to deviations from brownian motion. Systematic Biology 45: 27–47.
31. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290: 1151–1155. Available: <https://www.ncbi.nlm.nih.gov/pubmed/11073452>.
32. Pegueroles C, Laurie S, Albà MM (2013) Accelerated evolution after gene duplication: A time-dependent process affecting just one copy. Mol Biol Evol 30: 1830–1842. Available: <https://www.ncbi.nlm.nih.gov/pubmed/23625888>.
33. Han MV, Demuth JP, McGrath CL, Casola C, Hahn MW (2009) Adaptive evolution of young gene duplicates in mammals. Genome Res 19: 859–867. Available: <https://www.ncbi.nlm.nih.gov/pubmed/19411603>.
34. Cusack BP, Wolfe KH (2007) Not born equal: Increased rate asymmetry in relocated and retrotransposed rodent gene duplicates. Mol Biol Evol 24: 679–686. Available: <https://www.ncbi.nlm.nih.gov/pubmed/17179139>.
35. Scannell DR, Wolfe KH (2008) A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. Genome Res 18: 137–147. Available: <https://www.ncbi.nlm.nih.gov/pubmed/18025270>.
36. Pich I Roselló O, Kondrashov FA (2014) Long-term asymmetrical acceleration of protein evolution after gene duplication. Genome Biol Evol 6: 1949–1955. Available: <https://www.ncbi.nlm.nih.gov/pubmed/25070510>.
37. McGrath CL, Gout JF, Doak TG, Yanagi A, Lynch M (2014) Insights into three whole-genome duplications gleaned from the *paramecium caudatum* genome sequence. Genetics 197: 1417–1428. Available: <https://www.ncbi.nlm.nih.gov/pubmed/24840360>.

38. Purvis A, Rambaut A (1995) Comparative analysis by independent contrasts (caic): An apple macintosh application for analysing comparative data. *Comput Appl Biosci* 11: 247–251. Available: <https://www.ncbi.nlm.nih.gov/pubmed/7583692>.
39. Orme D (2018) The caper package: Comparative analysis of phylogenetics and evolution in r. Available: <https://cran.r-project.org/web/packages/caper/vignettes/caper.pdf>.
40. R Core Team (2018) R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Available: <https://www.R-project.org/>.
41. Kachroo AH, Laurent JM, Yellman CM, Meyer AG, Wilke CO, et al. (2015) Evolution. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. *Science* 348: 921–925. Available: <https://www.ncbi.nlm.nih.gov/pubmed/25999509>.
42. Laurent JM, Garge RK, Teufel AI, Wilke CO, Kachroo AH, et al. (2019) Humanization of yeast genes with multiple human orthologs reveals principles of functional divergence between paralogs. *bioRxiv* doi: <https://doi.org/10.1101/668335>.
43. Yu G, Smith D, Zhu H, Guan Y, Lam TT-Y (2017) Ggtree: An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* 8: 28–36. doi:[10.1111/2041-210X.12628](https://doi.org/10.1111/2041-210X.12628).
44. Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shkler M, et al. (2005) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* 21: 650–659. Available: <Go to ISI>://WOS:000227241200012.
45. Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, et al. (2011) The evolution of gene expression levels in mammalian organs. *Nature* 478: 343–348. Available: <https://www.ncbi.nlm.nih.gov/pubmed/22012392>.
46. Revell LJ (2012) Phytools: An r package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* 3: 217–223. doi:[doi:10.1111/j.2041-210X.2011.00169.x](https://doi.org/10.1111/j.2041-210X.2011.00169.x).
47. Guangchuang Y (2018) Treeio: Base classes and functions for phylogenetic tree input and output. Available: <https://guangchuangyu.github.io/software/treeio>.
48. Pennell MW, Eastman JM, Slater GJ, Brown JW, Uyeda JC, et al. (2014) Geiger v2.0: An expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics* 30: 2216–2218. Available: <https://www.ncbi.nlm.nih.gov/pubmed/24728855>.
49. Wickham H (2019) Stringr: Simple, consistent wrappers for common string operations. Available: <https://CRAN.R-project.org/package=stringr>.
50. Antoine Lucas DE with contributions by, Tuszynski J, Bengtsson H, Urbanek S, Frasca M, et al. (2018) Digest: Create compact hash digests of r objects. Available: <https://CRAN.R-project.org/package=digest>.
51. Wickham H, François R, Henry L, Müller K (2019) Dplyr: A grammar of data manipulation. Available: <https://CRAN.R-project.org/package=dplyr>.
52. Wickham H (2017) Tidyverse: Easily install and load the 'tidyverse'. Available: <https://CRAN.R-project.org/package=tidyverse>.
53. Słowiński K (2018) Ggrepel: Automatically position non-overlapping text labels with 'ggplot2'. Available: <https://CRAN.R-project.org/package=ggrepel>.
54. Warnes GR, Bolker B, Lumley T (2018) Gtools: Various r programming tools. Available: <https://CRAN.R-project.org/package=gtools>.
55. Wickham H (2016) Ggplot2: Elegant graphics for data analysis. Springer-Verlag New York. Available: <https://ggplot2.tidyverse.org>.
56. Wilke CO (2019) Cowplot: Streamlined plot theme and plot annotations for 'ggplot2'. Available: <https://CRAN.R-project.org/package=cowplot>.

57. Kassambara A (2014) EasyGgplot2: Perform and customize easily a plot with ggplot2. Available: <http://www.sthda.com>.
58. Auguie B (2017) GridExtra: Miscellaneous functions for "grid" graphics. Available: <https://CRAN.R-project.org/package=gridExtra>.
59. Urbanek S (2013) Png: Read and write png images. Available: <https://CRAN.R-project.org/package=png>.
60. Paradis E, Claude J, Strimmer K (2004) APE: Analyses of phylogenetics and evolution in r language. Bioinformatics 20: 289–290. Available: <https://www.ncbi.nlm.nih.gov/pubmed/14734327>.
61. Benjamini Y, Yekutieli D (2005) Quantitative trait loci analysis using the false discovery rate. Genetics 171: 783–790. Available: <https://www.ncbi.nlm.nih.gov/pubmed/15956674>.
62. Hochberg Y, Benjamini Y (1990) More powerful procedures for multiple significance testing. Stat Med 9: 811–818. Available: <https://www.ncbi.nlm.nih.gov/pubmed/2218183>.