Syst. Biol. 64(5):824–837, 2015

DOI:10.1093/sysbio/syv041
Advance Access publication June 22, 2015

Can We Identify Genes with Increased Phylogenetic Reliability?

VINSON P. DOYLE[1,2], RANDEE E. YOUNG[1,3], GAVIN J. P. NAYLOR[4], AND JEREMY M. BROWN[1,5,*]

[1]*Department of Biological Sciences and* [2]*Department of Plant Pathology and Crop Physiology, Louisiana State University, Baton Rouge, LA 70803, USA;* [3]*Department of Biology, University of Utah, Salt Lake City, UT 84112, USA;* [4]*Department of Biology and Hollings Marine Laboratory, College of Charleston, Charleston, SC 29424, USA;* [5]*Museum of Natural Science, Louisiana State University, Baton Rouge, LA 70803, USA*
[*]*Correspondence to be sent to: Department of Biological Sciences and Museum of Natural Science, Louisiana State University, 202 Life Science Building, Baton Rouge, LA 78703, USA; E-mail: jembrown@lsu.edu.*

Received 26 November 2014; reviews returned 27 January 2015; accepted 9 June 2015
Associate Editor: Vincent Savoleinen

Abstract.—Topological heterogeneity among gene trees is widely observed in phylogenomic analyses and some of this variation is likely caused by systematic error in gene tree estimation. Systematic error can be mitigated by improving models of sequence evolution to account for all evolutionary processes relevant to each gene or identifying those genes whose evolution best conforms to existing models. However, the best method for identifying such genes is not well established. Here, we ask if filtering genes according to their clock-likeness or posterior predictive effect size (PPES, an inference-based measure of model violation) improves phylogenetic reliability and congruence. We compared these approaches to each other, and to the common practice of filtering based on rate of evolution, using two different metrics. First, we compared gene-tree topologies to accepted reference topologies. Second, we examined topological similarity among gene trees in filtered sets. Our results suggest that filtering genes based on clock-likeness and PPES can yield a collection of genes with more reliable phylogenetic signal. For the two exemplar data sets we explored, from yeast and amniotes, clock-likeness and PPES outperformed rate-based filtering in both congruence and reliability. [molecular clock; phylogenomics; posterior prediction; rate of evolution; systematic error.]

Genome-scale data are increasingly being used for phylogenetic inference. One consistent finding in such studies is that topology varies considerably across genes. Much of this variation is to be expected as a consequence of incomplete lineage sorting and/or horizontal gene transfer (Pamilo and Nei 1988; Maddison 1997; Philippe and Douady 2003). However, some of this variation may result from using models that do not adequately account for the evolutionary processes that have shaped the patterns in the data. Although initial optimism held that increasing amounts of data would resolve difficult problems in phylogenetics (Gee 2003; Rokas et al. 2003), we are now finding that different genome-scale data sets frequently support contradictory hypotheses, each with strong support (e.g., Rokas et al. 2003; Phillips et al. 2004; Dunn et al. 2008; Philippe et al. 2009; Schierwater et al. 2009; Nosenko et al. 2013). The reason for this conflict is that increasing data set size may reduce stochastic error (Philippe et al. 2005, 2011), but it can also exacerbate systematic error leading to high confidence in the wrong tree (Kumar et al. 2012).

Two routes exist for reducing inferential artifacts due to systematic error: (i) improving phylogenetic models to better account for the complexity of molecular evolution (e.g., Yang 1994; Tillier and Collins 1995; Lartillot and Philippe 2004) or (ii) selecting subsets of the data that fit the available models (Philippe et al. 2005). The second option has historically been limited by data availability, but that constraint was lifted with the advent of high-throughput sequencing. The most common approach for selecting subsets of data is filtering based on the rate of evolution. Fast-evolving genes or sites are removed, as they have a tendency to produce complex character patterns that are not captured by simple models (e.g., Lopez et al. 2000; Nozaki et al. 2007). Despite the frequent use of rate, some have suggested that there are no "identifiable parameters" to indicate the most phylogenetically reliable genes (Gee 2003; Rokas et al. 2003; Salichos and Rokas 2013). However, we contend that many approaches for selecting preferred genes or sites remain un- or underinvestigated.

In this study, we investigate the performance of two underutilized approaches to phylogenomic filtering, and compare them to the use of rate. Our intention is not to exhaustively explore all approaches, but simply to ask whether "identifiable parameters" for filtering might exist. Our first approach is to prefer genes that evolve in a clock-like manner, as might be expected when molecular constraints, population sizes, and selection pressures remain constant over time and across lineages (Kimura 1964, 1968). Our motivation for focusing on clock-like genes is to avoid long-branch-attraction artifacts (Felsenstein 1978; Brinkmann et al. 2005; Zhong et al. 2011), and because clock-like genes have desirable distance properties. Notably, they remain clock-like even when distances are estimated using an incorrect model (Steel and Penny 2000). Under this model, we expect that analyses of clock-like genes are less likely to result in systematically biased inferences. Our focus on the degree to which a gene deviates from the molecular clock is intended to serve as a proxy for the potential of model misspecification to result in biased inferences. We are not interested in deviations from the clock for their own sake, and we expect that if all models are adequate, clock-like genes should be no more reliable than nonclock-like genes.

Our second approach to filtering tests fit between the model of sequence evolution and the data using Bayesian posterior prediction (Bollback 2002; Brown 2014). Models of sequence evolution that adequately

capture the evolutionary processes relevant to the data should produce the most reliable topological inferences. Although comparisons of relative model fit are commonplace in phylogenetics (i.e., model choice), evaluation of absolute fit between the data and the selected model is not. The latter can be accomplished in a Bayesian posterior predictive framework by simulating data sets using trees and parameter values drawn from the posterior distribution (Bollback 2002). The simulated data, also known as the posterior predictive distribution, can then be compared with the empirical data. If the model of sequence evolution fits the empirical data well, the empirical data set should be a plausible draw from the posterior predictive distribution. In this study, we compare the phylogenetic information contained in the empirical and simulated data using test statistics introduced by Brown (2014).

If some of the topological variation across genes is driven by systematic error and our approaches are able to distinguish between genes based on their susceptibility to such error, we expect to see two patterns. First, genes judged to be more reliable (by rate, clock-likeness, or posterior prediction) should be more similar to well-established reference topologies, when they are known. Second, such reliable genes would also be expected, in many circumstances, to have more congruent topologies. We applied these three filtering approaches to two empirical phylogenomic data sets where such reference topologies exist and tested the efficacy of the filtering based on the two criteria outlined above.

Methods

Multiple Sequence Alignments

We evaluated the impact of the three different filtering approaches on gene tree reliability using two previously published data sets (Hess and Goldman 2011; Crawford et al. 2012). The first includes nucleotide sequences of 343 protein-coding orthologs from 18 yeast species in the phylum Ascomycota, subphylum Saccharomycotina. We briefly outline the methods used by Hess and Goldman 2011 to assemble these multiple sequence alignments (MSAs), but see their paper for additional details. Hess and Goldman (2011) collected 1148 orthologous genes for 14 yeast species included in the Fungal Orthogroups Repository (FOR) at the Broad Institute (Wapinski et al. 2007). After applying a series of stringent filtering steps to reduce the inclusion of paralogous loci from the remaining 4 species not present in FOR, 343 genes with amino acid sequence data for all 18 species were aligned using MAFFT v. 6.24 (Katoh and Toh 2008) and uncertain regions of the alignment were trimmed using Gblocks v. 0.91b (Castresana 2000). The corresponding nucleotide alignments were reconstructed using BLAT (Kent 2002). The minimum, maximum, mean, and median length of the yeast MSAs was 264, 3435, 943, and 837 nucleotides, respectively.

The second set of MSAs consists of 1145 ultraconserved elements (UCEs) from 10 amniote species assembled by Crawford et al. (2012) using previously published genome sequences and *de novo* genomic enrichment (see Crawford et al. 2012 and Faircloth et al. 2012 for details). The amniote UCE sequences were aligned with MUSCLE (Edgar 2004). Loci missing nucleotide sequence data from any taxa were excluded. The minimum, maximum, mean, and median length of the amniote UCE MSAs was 129, 741, 406, and 403 nucleotides, respectively.

Model Selection and Maximum-Likelihood Gene Tree Estimation

We selected the best-fit model of nucleotide sequence evolution for each locus from a set of 24 models using Akaike's Information Criterion (AIC; Akaike 1974), as implemented in MrModelTest 2.3 (Nylander 2004) and PAUP* v4b10 (Swofford 2003). We inferred the maximum-likelihood (ML) phylogeny for each locus assuming the AIC-selected model in Garli v2.0 (Zwickl 2006), using five replicate searches. Branches of near-zero length ($\leqslant 1 \times 10^{-8}$) were collapsed to create polytomies. Each search was terminated after 5000 generations without an improvement of 0.01 or more log-likelihood units. Branch lengths for the reference topology of Hess and Goldman (2011) are ML estimates based on a concatenated alignment of all 343 genes assuming a GTR + I +Γ model of sequence evolution.

Molecular Clock and Evolutionary Rate Filtering

Given the selected model of sequence evolution and the ML tree for each gene, we calculated likelihoods twice: once enforcing a strict clock model and once estimating each branch length independently. In both cases, likelihoods were calculated in PAUP* v4b10 (Swofford 2003) using parameter and branch-length (in the case of heterogeneous rates) estimates from Garli v2.0 (Zwickl 2006). The likelihood ratio between these models was calculated as twice the difference in their log-likelihood scores and used as a measure of clock-likeness for each gene. Genes were then sorted in ascending order by their clock-likeness (likelihood ratios) and binned into deciles such that each bin contained 34 genes for the yeast data and 114 genes for the amniote data. We also compared the impact of two different rooting schemes, midpoint and outgroup rooting, on the likelihood ratio and subsequent binning. Results from these two different approaches to rooting were largely congruent, so hereafter we focus on analyses using midpoint rooting. For the sake of completeness, we also examined the frequency with which genes in each data set rejected a molecular clock using a chi-squared test (Felsenstein 1981), although we are not primarily interested in testing the molecular clock hypothesis. Our objective was to use the likelihood ratios from a clock test as a relative measure of clock-likeness across genes. The

relative evolutionary rate of each gene was scored as the sum of the branch lengths on its ML tree. As with clock-likeness, we sorted genes according to evolutionary rate in ascending order and then binned them into deciles.

Posterior Predictive Filtering

Posterior prediction is a Bayesian statistical procedure for checking the fit of a model to the data being analyzed. Briefly, posterior prediction in a phylogenetic analysis involves (i) drawing trees and parameters values from a posterior distribution, (ii) using them to simulate new data sets (known as posterior predictive data sets), (iii) summarizing each data set using a relevant test statistic, and then (iv) comparing the empirical test statistic value to the simulated distribution. Different test statistics may be employed to assess different aspects of fit. Here, we employ a combination of statistics that aim to capture model violations affecting topological inferences (Brown 2014).

More specifically, we estimated the joint posterior distribution of parameter values and tree topologies for each empirical data set using MrBayes v3.2.1 (Ronquist and Huelsenbeck 2003; Ronquist et al. 2012) with 4 replicate Markov chain Monte Carlo (MCMC) runs and 4 Metropolis-coupled chains per run, assuming the same model of sequence evolution as used for ML estimation. Each run consisted of 5 million generations with a sampling frequency of 2000 for yeast data, and 2.5 million generations with a sampling frequency of 1000 for amniote UCE data. Convergence checking was done using the approach outlined by Brown and Lemmon (2007) as implemented in MrConverge v1b2.5 (Lemmon 2007).

One hundred tree topologies and associated parameter values were sampled from the posterior distribution for each gene by drawing 25 samples evenly spaced across generations from the stationary distribution of each replicate run. A separate posterior predictive data set was then simulated for each sample with PuMA v0.905 (Brown and ElDabaje 2009), which relies on Seq-Gen (Rambaut and Grassly 1997). Because Seq-Gen does not simulate MSAs with missing data, we substituted missing data for nucleotides in each posterior predictive alignment to match the patterns of missing data observed in the empirical data using a custom python script (available from http://github.com/jembrown/repMissPatterns and http://dx.doi.org/10.5061/dryad.fd3m4). Each of these simulated posterior predictive data sets was then analyzed with MrBayes v3.2.1 in the same manner as the empirical data, but reducing the number of generations to 2 million and 1 million for yeast and UCE data, respectively. In general, convergence occurs more quickly when analyzing posterior predictive data sets, due to the close matching between generating and assumed models of sequence evolution. Figure 1, replicated from Brown (2014), provides a schematic

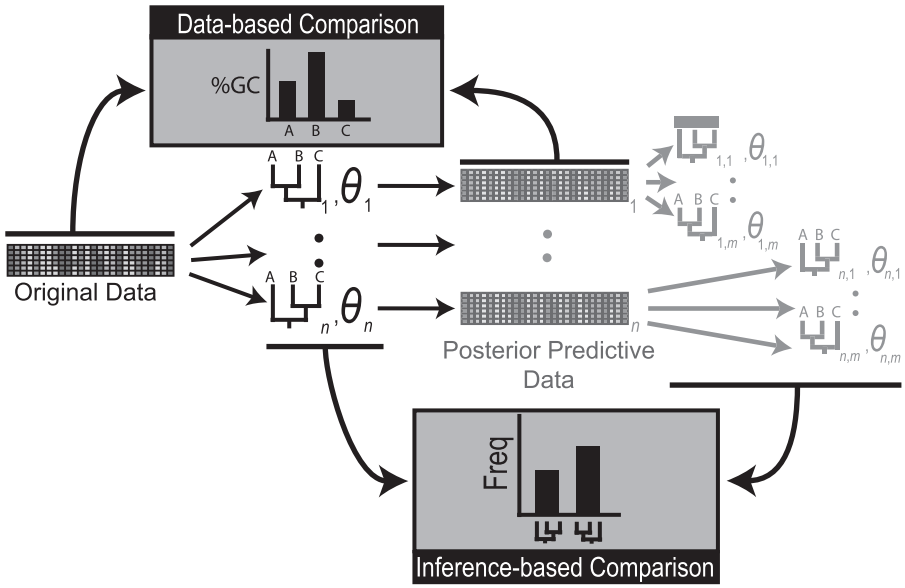

FIGURE 1. A schematic representation of data- and inference-based approaches to assessing model plausibility using posterior predictive simulation. Most statistics proposed for testing model plausibility compare data-based characteristics of the original data set to the posterior predictive data sets (e.g., variation in GC-content across species). The approach used in this study utilizes test statistics that compare the inferences resulting from different data sets (e.g., the distribution of posterior probability across topologies). MSAs are represented as shaded matrices and arrows originating from MSAs point to the MCMC samples of tree topologies and scalar model parameters (θ) resulting from Bayesian analysis of that MSA. Subscripts of MCMC samples taken during analysis of the original data index the samples (1, …, n). Subscripts for each posterior predictive data set indicate which MCMC sample was used in its simulation. Subscripts for MCMC samples resulting from analysis of a posterior predictive data set first indicate the posterior predictive data set that was analyzed and next index the MCMC samples from analysis of that particular data set (1, …, m). This figure is reproduced from Brown (2014).

representation of inference-based posterior predictive assessment of model fit.

To detect disparities between the inferences drawn from analyses of empirical data and those drawn from posterior predictive data, we calculated several test statistics that summarize relevant aspects of each posterior distribution using AMP v0.99e (Brown 2014). Because we are interested in detecting model violations that specifically impact topological inferences, we focus on two classes of test statistics that summarize marginal distributions of topologies: the distribution of symmetric differences (unweighted Robinson–Foulds; Robinson and Foulds 1981) among trees in the posterior and the difference in statistical entropy between the prior and posterior distributions of topologies (see Brown 2014 for more details). We summarized all postburnin samples (m samples) from each posterior distribution by examining several quantiles in the ordered vector of pairwise symmetric differences (of length $\binom{m}{2}$), including the 1st quartile, median, 3rd quartile, 99th percentile, 999th 1000 quantile, and the 9999th 10,000 quantile. Each of these quantiles is a potential test statistic. The entropy test statistic measures the change in the distribution of probability across topologies when comparing the prior to the posterior. This quantity can be interpreted as the gain in information provided by the data, conditional on the assumed model (again, see Brown 2014 for more details). A uniform distribution across topologies has maximum entropy, whereas a single topology with a posterior probability of 1 has minimum entropy. Figure 2 in Brown (2014) provides example topological test statistic calculations from a posterior distribution.

To quantify the position of the empirical value relative to each posterior predictive distribution, we calculated effect sizes. We chose to use effect sizes rather than P-values to differentiate between empirical values that lie near, but outside, the distribution of posterior predictive values from those that are very far outside. Effect size was calculated as the absolute value of the difference between the median posterior predictive value and the empirical value divided by the standard deviation of the posterior predictive distribution. We calculated effect sizes in the same manner for all test statistics.

In order to summarize topological model fit with a single value, we calculated the mean effect size from one quantile-based test statistic and the entropy test statistic. We selected the single quantile-based test statistic with the fewest effect sizes of zero for each data set in order to maximize our power to detect model violation. We chose a single quantile-based test statistic so as not to unduly weight the mean effect size toward this class of statistics. Genes were then sorted in ascending order by their mean effect sizes and binned into deciles such that each bin contained 34 genes for the yeast data and 114 genes for the amniote data. Hereafter, we use PPES to refer to posterior predictive effect sizes.

Tree Distances

We assessed the utility of the three filtering approaches described above using two criteria based on tree-to-tree distances. First, we determined if there was a significant reduction in the mean distance to the reference topology for trees in each decile. Sets of trees equal in number to the size of each decile (34 for the yeast orthologs and 114 for the amniote UCEs) were randomly sampled from the full set, the mean distance to the reference topology was calculated, and this procedure was repeated 100,000 times. Reductions in distances to the reference topology were considered significant if they were in the lower 5% of the null distribution. We also calculated the Spearman's rank correlation coefficient, r_s, between the decile indices and the ranked mean pairwise distances and assessed significance with the exact one-tailed P-value, using algorithm AS 89 (Best and Roberts 1975), as implemented in the "stats" package in R v2.15.1 (R Core Development Team 2012). Second, we determined if there was a significant reduction in pairwise distances among ML trees in the lowest deciles of ranked sets and/or a significant increase in pairwise distances among trees in the upper deciles of ranked sets by calculating the mean pairwise distance between ML trees within each decile and comparing that value to expectations based on random sampling. Similar to the distance-to-reference comparisons, the null distribution was generated by randomly sampling the same number of ML trees as found in each decile (34 for the yeast orthologs and 114 for the amniote UCEs) from the full set of ML trees, calculating the mean pairwise distance, and repeating this procedure 100,000 times. Reductions in pairwise distances were considered significant if they were in the lower 5% of the null distribution. Plots showing intradecile comparisons and corresponding null distributions are presented with the Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.fd3m4.

Several metrics have been proposed to measure the distance between two trees in tree space with the aim of summarizing their topological similarity or dissimilarity. The symmetric difference between two trees is a commonly used metric, but some of its properties are undesirable for tree comparison. First, the rearrangement of a single taxon can maximize the symmetric difference between two trees despite partial similarities between them (Penny et al. 1982; Felsenstein 2004; Lin et al. 2012). Second, the distribution of distances between two random binary trees is strongly skewed in the direction of larger distances, reducing the power to discriminate between sets of trees based on similarity (Penny et al. 1982; Lin et al. 2012). Another metric, the matching distance, was recently introduced and has a broader distribution of distances between random trees. The matching distance provides more discriminatory power and is not as sensitive to minor leaf rearrangements as the symmetric difference (Bogdanowicz and Giaro 2012; Bogdanowicz et al. 2012; Lin et al. 2012). We characterized

the similarity of trees using both the symmetric difference, because of its familiarity, and the matching distance (both matching split and matching cluster), because of its discriminatory power and robustness to minor topological changes. Symmetric differences were calculated with the Dendropy Phylogenetic Computing Library v3.12.0 (Sukumaran and Holder 2010) and matching distances were calculated using TreeCmp v1.0-b291 (Bogdanowicz et al. 2012). We used the matching split metric, a measure designed for calculating distances among unrooted phylogenetic trees (Bogdanowicz and Giaro 2012), to calculate distances among the inferred ML trees. In order to calculate distances between ML trees and reference topologies, we used the matching cluster distance, which is analogous to the matching split distance but designed for comparisons of rooted topologies (Bogdanowicz and Giaro 2013). Results based on symmetric differences are presented in the Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.fd3m4.

For comparisons with reference trees, we also used a third criterion to assess the utility of the filtering approaches: the number of splits in each ML topology that conflict with the reference topology. This criterion is closely related to the symmetric difference, but focuses exclusively on "false" splits in the ML trees. When both the ML tree and the reference tree are fully bifurcating, this value is simply half the symmetric difference. However, this relationship does not hold when either the reference tree or the gene tree contains polytomies. We assessed the significance of these values in the same manner as the tree-to-tree distances, generating a null distribution based on 100,000 resamplings. We also calculated r_s as above and assessed significance of the correlation using a one-tailed P-value. Plots of the null distributions and the mean number of conflicting splits within each decile are presented with the Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.fd3m4.

Conflicting Splits

In addition to characterizing overall topological similarity, we also examined split-specific conflict between ML trees and the corresponding reference topology by calculating the number of trees in each decile that conflict with each split in the reference topology. To look for significant increases or decreases in split-specific conflict, we compared these values for each decile to expectations based on a null distribution generated by 10,000 random resamplings. We specifically tested for significant (P <0.05) reductions in conflict in the lowest deciles and significant increases in conflict in the upper deciles with and without adjusting for multiple comparisons using the Bonferroni and false discovery rate (FDR) (Benjamini and Hochberg 1995) correction methods, with the number of comparisons equal to twice the number of splits in each reference topology.

Results

Molecular Clock, Evolutionary Rate, and Posterior Predictive Filtering

The distribution of test statistics from the molecular clock, evolutionary rate, and posterior predictive tests are summarized in Table S1 and Fig. S1 available as the Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.fd3m4. LRT values for clock-likeness varied from 7.27 to 933.21 for the yeast data with a median of 144.5, and from 2.08 to 163.9 for the amniote UCE data with a median of 29.71. The 999th 1000 quantile was the posterior predictive quantile-based metric with the fewest effect sizes of zero for the yeast data, giving an effect size of zero for only 12 yeast orthologs, whereas the quantile-based metric with the fewest effect sizes of zero for the amniote UCE data was the 99th percentile. The distribution of tree lengths is nonoverlapping between the yeast ortholog and amniote UCE data sets (Table S1; Fig. S1e,f available as the Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.fd3m4).

Distance between Deciles and Reference Topologies

Yeast orthologs.—The mean matching distances between the ML trees and the yeast reference topology in four of the five lowest (most clock-like) deciles of likelihood ratio test statistics are significantly smaller than expected (Fig. 2a). Yeast ortholog deciles also exhibit a strong positive correlation ($r_s = 0.952$, $P = 2.2 \times 10^{-16}$) between clock-likeness ranks and the ranked mean matching distance to the reference tree. In addition, the mean number of splits in each ML tree that conflict with the reference topology is significantly reduced in the first, second, and third deciles when yeast ortholog ML trees are ranked by clock-likeness (Fig. S2a available as the Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.fd3m4). The rank correlation across deciles between clock-likeness and the mean number of splits that conflict with the reference is very strong ($r_s = 0.957$, $P = 6.9 \times 10^{-6}$).

Ranking ML trees by PPES does not yield individual deciles with significantly smaller mean matching distances to the yeast reference topology (Fig. 2b). The rank correlation between PPES and mean distance to the reference tree is much weaker than observed for clock-based filtering ($r_s = 0.394$, $P = 0.1314$). Ranking according to PPES does lead to a significant reduction in the mean number of splits in conflict with the reference tree for the first and second deciles, but the overall relationship between conflict and PPES is not as strong (Fig. S2b available as the Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.fd3m4; $r_s = 0.600$, $P = 0.03656$) as it was for clock-likeness.

Similar to ranking ML trees by PPES, ranking by rate of evolution does not yield individual deciles with significantly smaller mean distances to the reference topology (Fig. 2c). The rank correlation

between rate of evolution and ranked mean distance to the reference tree is moderately negative and significant ($r_S = -0.576$, $P = 0.04388$). Ranking by rate of evolution does lead to a significant reduction in the mean number of conflicting splits for the fourth, fifth, and eighth deciles, but not the lowest deciles (Fig. S2c available as the Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.fd3m4). Overall, the rank correlation between rate of evolution and conflict with the reference is weak (Fig. S2c available as the Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.fd3m4; $r_S = 0.103$, $P = 0.3925$).

Amniote UCEs.—The mean matching distances between the amniote reference topology and amniote UCE ML trees in the third and sixth clock-likeness deciles are significantly smaller than expected (Fig. 3a). The mean distances for each of the lower six deciles are smaller than the median distance in the null distribution, whereas the mean distance for each of the upper four deciles are larger than the median distance. Across deciles, clock-likeness ranks and ranks in mean distance to the reference topology exhibit a moderately strong correlation (Fig. 3a; $r_S = 0.709$, $P = 0.01376$). Ranking by clock-likeness leads to a significant reduction in the number of splits that conflict with the reference for the third decile (Fig. S3a available as the Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.fd3m4), with a moderate rank correlation between clock-likeness and conflict across all deciles ($r_S = 0.711$, $P = 0.01055$).

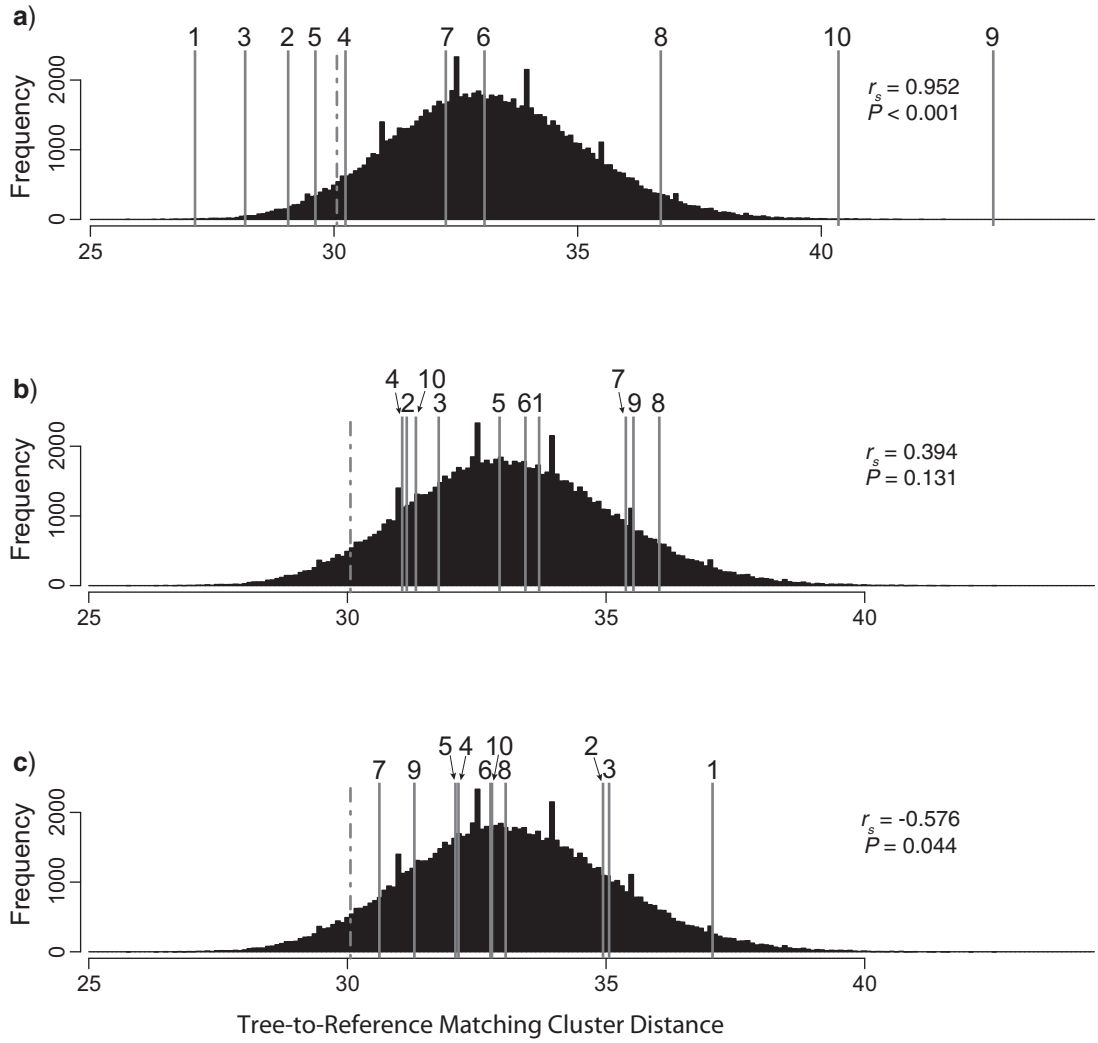

FIGURE 2. Mean matching cluster distance between yeast ortholog ML trees and the reference topology for each filtered decile. a) Mean matching cluster distance to the reference topology within each decile ranked by likelihood ratio test statistics. b) Mean matching cluster distance to the reference topology within each decile ranked by PPES. c) Mean matching cluster distance to the reference topology within each decile ranked by rate of evolution. The null distribution was estimated by resampling 34 (10%) ML yeast ortholog trees and recalculating the mean matching split distance to the reference topology 100,000 times. r_s is Spearman's rank correlation coefficient between decile indices and the ranked mean distances to the reference topology.

The ML trees in the lowest three deciles of PPES are significantly more similar to the reference topology than expected, based on matching distances (Fig. 3b). The mean distances for each of the lowest five deciles are all smaller than the median distance in the null distribution and the mean distances for each of the upper five deciles are larger than the median. In addition, there is a very strong rank correlation (Fig. 3b; $r_S = 0.985$, $P = 1.144 \times 10^{-7}$) between PPES and mean distance to the reference topology. Posterior predictive ranking also yields a significant reduction in the mean number of splits in conflict with the reference tree for the lowest two deciles (Fig. S3b available as the Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.fd3m4), and the third decile is marginally significant ($0.042 \leqslant P \geqslant 0.052$). As with matching distances, the rank correlation between PPES and conflict with the reference is very strong (Fig. S3b available as the Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.fd3m4; $r_S = 0.964$, $P = 2.2 \times 10^{-16}$).

In contrast to rankings based on clock-likeness or PPES, rankings based on rate of evolution do not yield trees that have smaller matching distances to the reference topology in the lowest ranked deciles (Fig. 3c). In fact, the deciles containing the trees with the fastest relative rate of evolution (highest ranks) have the smallest mean distances to the reference topology. The mean distance in deciles eight and ten is significantly smaller than expected and each of the upper five deciles are smaller than the median distance in the null distribution. Correspondingly, each of the lower five deciles has a mean distance larger than the median value in the null. This distribution of ranks leads to a very strong negative rank correlation (Fig. 3c; $r_S = -0.939$, $P = 2.2 \times 10^{-16}$) between rate of evolution and mean matching distance to the

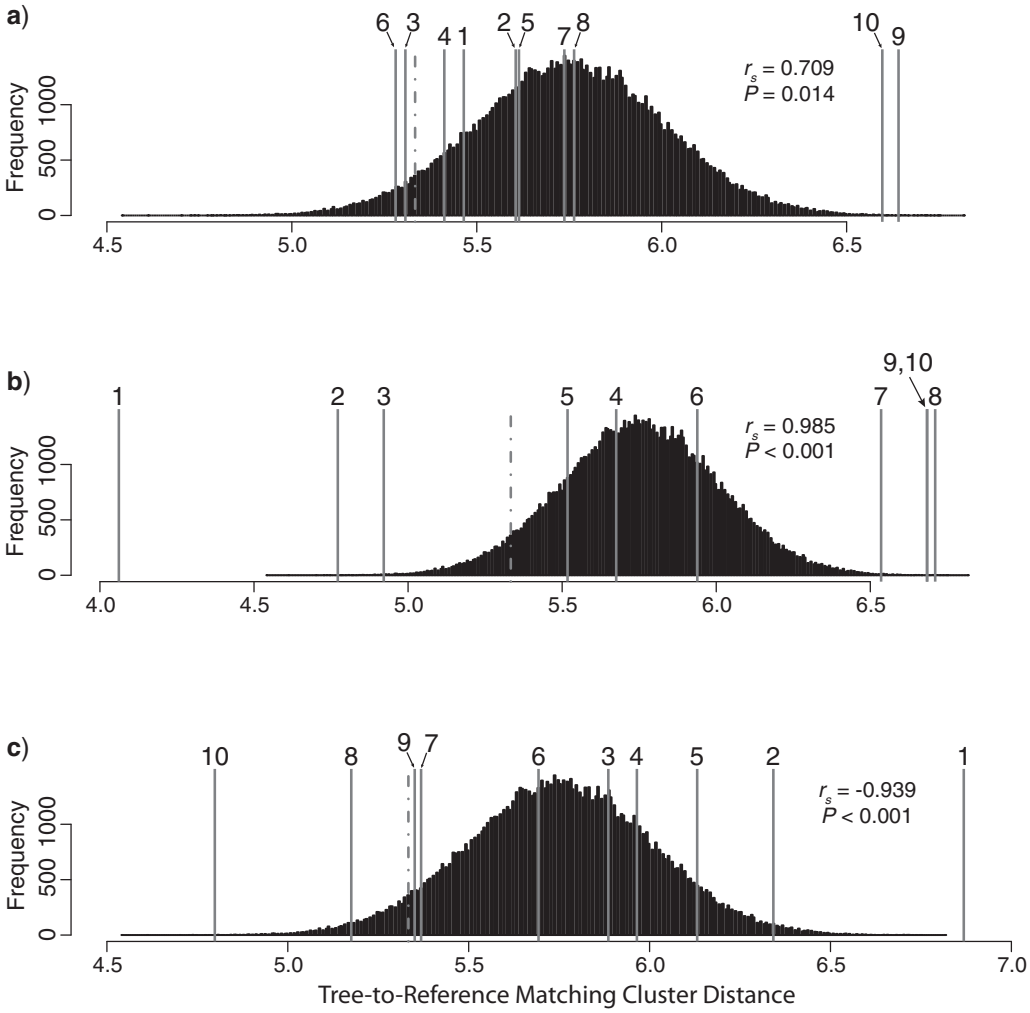

FIGURE 3. Mean matching cluster distance between amniote UCE ML trees and the reference topology for each filtered decile. A) Mean matching cluster distance to the reference topology within each decile ranked by likelihood ratio test statistics. B) Mean matching cluster distance to the reference topology within each decile ranked by PPES. C) Mean matching cluster distance to the reference topology within each decile ranked by rate of evolution. The null distribution was estimated by resampling 114 (10%) ML amniote UCE trees and recalculating the mean matching split distance to the reference topology 100,000 times. Numbers separated by commas indicate means that are identical. r_S is Spearman's rank correlation coefficient between decile indices and the ranked mean distances to the reference topology.

TABLE 1. Number of trees in different deciles conflicting with each split on the yeast reference topology depicted in Figure 4

Split	Total (%)	Most Clock-like	Least Clock-like	Lowest PPES	Highest PPES	Slowest	Fastest
A	54 (15.7)	**0 (0)[a]**	**14 (41.2)[a]**	6 (17.6)	2 (5.9)	6 (17.6)	**17 (50)[a]**
B	70 (20.4)	**1 (2.9)[a]**	**17 (50)[a]**	4 (11.8)	6 (17.6)	9 (26.5)	**19 (55.9)[a]**
C	77 (22.4)	8 (23.5)	7(20.6)	7 (20.6)	10 (29.4)	6 (17.6)	10 (29.4)
D	147 (42.9)	15 (44.1)	**21 (61.8)[b]**	11 (32.4)	13 (38.2)	10 (29.4)	**28 (82.4)[a]**
E	36 (10.5)	**0 (0)**	6 (17.7)	3 (8.8)	3 (8.8)	4 (11.8)	4 (11.8)
F	32 (9.3)	1 (2.9)	**10 (29.4[a]**	1 (2.9)	2 (5.9)	**0 (0)**	**23 (67.6)[a]**
G	166 (48.4)	17 (50)	18 (52.9)	13 (38.2)	12 (35.3)	**9 (26.5)[b]**	**26 (76.5)[a]**
H	92 (26.8)	7 (20.6)	8 (23.5)	7 (20.6)	8 (23.5)	9 (26.5)	9 (26.5)
I	58 (16.9)	4(11.8)	6 (17.7)	2 (5.9)	7 (20.6)	7 (20.6)	7 (20.6)

Notes: Significant decreases/increases in split-specific conflict between ML trees and the corresponding reference topology are indicated by bold font. Values in parentheses are percentages of trees. $P-$values were estimated by comparison with a null distribution of split-specific conflict generated by 10,000 random resamplings.

[a]Significant after both Bonferroni and FDR correction.

[b]Significant after FDR correction only.

reference topology. Similar results hold when using the mean number of splits conflicting with the reference. Ranking by rate of evolution does not produce any deciles with significant reductions in conflict (Fig. S3c available as the Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.fd3m4), although the slowest evolving genes (decile 1) do show the least amount of conflict. Across all deciles, the rank correlation between rate of evolution and conflict is moderate (Fig. S3c available as the Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.fd3m4; $r_s = 0.482$, $P = 0.0793$).

Topological Similarity within Deciles

Yeast orthologs.—Topological similarity among trees inferred from yeast orthologs is significantly greater than expected in the deciles that are most clock-like, produce the most plausible inferences, and are the slowest evolving (Fig. S4 available as the Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.fd3m4). However, the correlation between decile indices and topological congruence varied across metrics. Yeast orthologs exhibit a strong positive correlation (Fig. S4a available as the Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.fd3m4; $r_s = 0.903$, $P = 0.00044$) between clock-likeness ranks and the ranked mean distance among trees in each decile, indicating that the most clock-like orthologs are also more similar topologically. Rankings based on PPES and rate of evolution are both moderately correlated with the ranked mean distance among trees in each decile, but this correlation was nonsignificant for rate (Fig. S4b,c available as the Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.fd3m4).

Amniote UCEs.—The 10% of most clock-like ML trees inferred from amniote UCEs are more topologically similar than expected and the lowest five deciles (and the seventh decile) have a mean pairwise distance smaller than the median (Fig. S5a available as the Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.fd3m4). Filtering by PPES significantly reduces topological heterogeneity among trees in four of the five lowest deciles (Fig. S5b available as the Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.fd3m4). Similarly, the decile composed of the shortest trees (genes with the slowest rates) has the smallest mean distance. Again, however, these metrics vary in their overall rank correlations. Although there is a strong positive correlation between PPES and topological congruence (Fig. S5b available as the Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.fd3m4; $r_s = 0.976$, $P = 2.2 \times 10^{-16}$), and clock-likeness and topological congruence (Fig. S5a available as the Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.fd3m4; $r_s = 0.891$, $P = 0.0006901$), there is a very weak negative rank correlation (Fig. S5c available as the Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.fd3m4; $r_s = -0.103$, $P = 0.6206$) between evolutionary rate and topological congruence.

Split-specific Conflict with Reference Topologies

Yeast orthologs.—The most clock-like decile contains significantly fewer trees in conflict with three splits on the yeast reference tree (Table 1; Fig. 4), while there is no significant change with respect to the remaining splits. The reduction in the number of trees in conflict with each split remains significant for two of these after both Bonferroni and FDR corrections (Benjamini and Hochberg 1995). In contrast to reductions in conflict in the lowest (most clock-like) decile, trees in the uppermost (least clock-like) decile exhibit a significant increase in conflict with four splits on the reference tree, all of which remain significant after Bonferroni and/or FDR adjustment (Table 1; Fig. 4).

Unlike clock-based filtering, trees in the lowest decile of PPES do not show a significant decrease

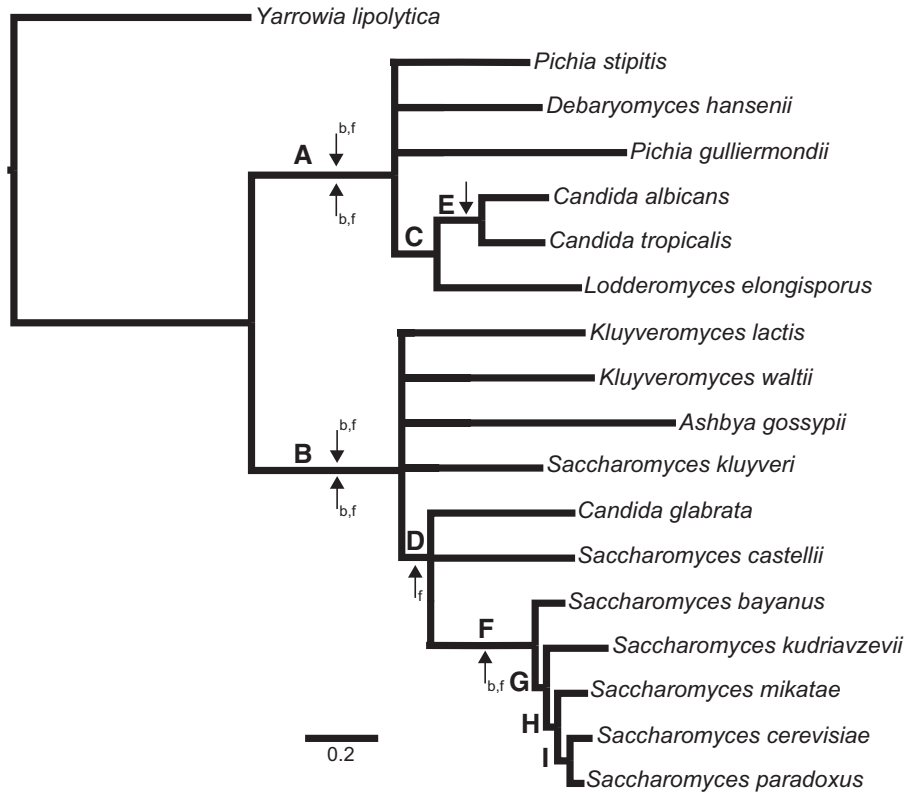

FIGURE 4. Split-specific changes in the percentage of trees that conflict with the yeast reference topology among the lower and upper deciles of clock-likeness. Downward arrows indicate splits with a significant reduction in conflict for trees in the most clock-like decile (lowest likelihood ratio test statistics). Upward arrows indicate splits with a significant increase in conflict for trees in the least clock-like decile (highest likelihood ratio test statistics). Those changes that remain significant after adjusting for multiple comparisons with Bonferroni and FDR are annotated with "b" and "f", respectively. Branch lengths are ML estimates from a concatenated alignment of all 343 yeast orthologs.

in conflict with any splits on the reference tree (Table 1). However, trees in this decile do exhibit marginally significant reductions in conflict with split I, as well as split F ($0.0117 \leqslant P \leqslant 0.0501$ and $0.0311 \leqslant P \leqslant 0.1542$, respectively), but do not remain so after either Bonferroni or FDR adjustment. Similarly, trees in the upper decile have no significant increases in conflict with any splits in the reference tree (Table 1). Trees in the slowest decile exhibit significantly reduced conflict with two splits on the reference tree, one of which remains significant after FDR adjustment (Table 1). Trees in the fastest decile exhibit a significant increase in conflict with five splits on the reference tree and these increases remain significant after both Bonferroni and FDR adjustment (Table 1).

Amniote UCEs.—Significantly fewer trees in the most clock-like decile are in conflict with two splits on the amniote reference tree (Table 2) and both reductions remain significant after Bonferroni and/or FDR adjustment. In addition, clock-like genes have a marginally significant ($0.0329 \leqslant P \geqslant 0.0585$) reduction in conflict with the split uniting birds (C in Fig. 5), before adjusting for multiple comparisons. Significantly more trees in the least clock-like (uppermost) decile conflict with the same two splits on the reference tree where we observed a significant decrease among the most clock-like genes, and both remain significant after adjusting for multiple comparisons with both Bonferonni and FDR.

Posterior predictive filtering significantly reduces the percentage of trees in conflict with all 5 nontrivial splits in the amniote reference tree (Table 2; Fig. 5) for the lowest decile, all of which remain significant after Bonferroni and/or FDR adjustment. The percentage of trees in the uppermost posterior predictive decile that conflict with reference splits is significantly increased for four of the five nontrivial splits on the reference tree, even after adjustments with Bonferroni and/or FDR. The only split for which there is not a significant increase in conflict is that uniting squamates (Table 2; E in Fig. 5).

In contrast to filtering by clock-likeness and PPES, filtering by rate did not yield a significant decrease in conflict with any split in the amniote reference tree for the slowest decile nor did it produce any significant increases in conflict for the fastest decile (Table 2).

DISCUSSION

Methods for quantifying the extent to which systematic error drives topological variation across gene trees, and for identifying those genes that are least

TABLE 2. Number of trees in different deciles conflicting with each split on the amniote reference topology depicted in Figure 5

Split	Total (%)	Most Clock-like	Least Clock-like	Lowest PPES	Highest PPES	Slowest	Fastest
A	625 (54.6)	**51 (44.7)[b]**	**77 (67.5)[a]**	**47 (41.2)[a]**	**78 (68.4)[a]**	63 (55.3)	60 (52.6)
B	606 (53.0)	74 (64.9)	62 (54.4)	**42 (36.8)[a]**	**74 64.9)[a]**	56 (49.1)	61 (53.5)
C	206 (18.0)	14 (12.3)	21 (18.4)	**7 (6.1)[a]**	**33 (28.9)[a]**	28 (24.6)	8 (7.0)
D	109 (9.5)	**2 (1.8)[a]**	**25 (21.9)[a]**	**1 (0.9)[a]**	**18 (15.8)[b]**	13 (11.4)	10 (8.8)
E	97 (8.5)	8 (7.0)	11 (9.6)	**3 (2.6)[b]**	12 (10.5)	11 (9.6)	11 (9.6)

Notes: Significant decreases/increases in split-specific conflict between ML trees and the corresponding reference topology are indicated by bold font. Values in parentheses are percentages of trees. *P*−values were estimated by comparison with a null distribution of split-specific conflict generated by 10,000 random resamplings.
[a]Significant after both Bonferroni and FDR correction.
[b]Significant after FDR correction only.

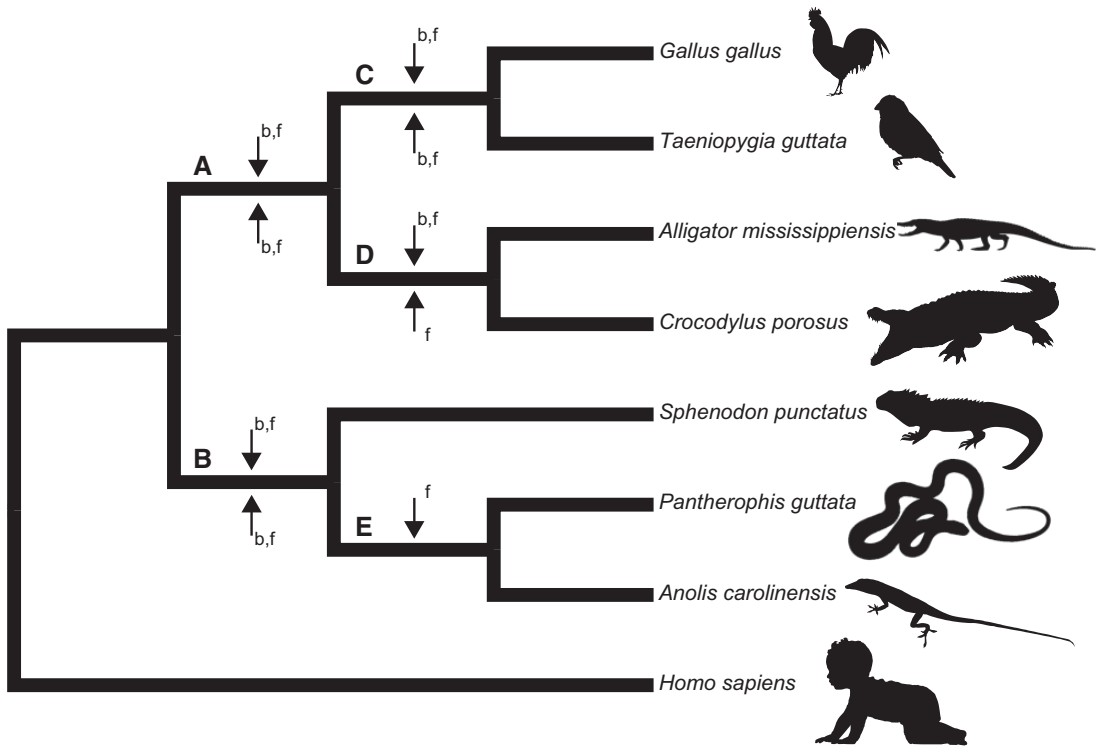

FIGURE 5. Split-specific changes in the percentage of trees that conflict with the amniote reference topology among the lower and upper deciles of PPES. Downward arrows indicate splits with a significant reduction in conflict for trees in the lowest decile of PPES. Upward arrows indicate splits with a significant increase in conflict for trees in the upper decile of PPES. Those changes that remain significant after adjusting for multiple comparisons with Bonferroni and FDR are annotated with "b" and "f", respectively.

susceptible to such error, are underdeveloped. Our results demonstrate that two little used approaches for ranking genes can increase gene tree reliability and congruence, although their relative performance varies across data sets. For two exemplar phylogenomic data sets with well-established reference trees, one of which is partially resolved, genes assessed to be more clock-like or with better fit to assumed models of sequence evolution exhibited greater topological similarity to reference trees and to each other. Both of these ranking schemes preferred genes with more reliable and congruent topologies than did the commonly used approach of ranking based on the rate of evolution.

Precise *a priori* expectations about the cause and prevalence of systematic bias in any particular data set are usually difficult to formulate. However, strong empirical evidence supports the influence of such bias across many data sets (e.g., Philippe et al. 2011). Identifying general features of genes that suggest their evolution matches model assumptions well, or at least does not violate model assumptions too severely, is one way to avoid the effects of systematic bias. Rate of evolution has been used as a common default proxy for this purpose, based on both theoretical considerations (e.g., Felsenstein 1978) and empirical experience (Lartillot and Philippe 2008; Chiari et al.

2012). However, many other features of genes have been relatively little explored or underutilized. Clock-like evolution, or lack thereof, has been characterized across many data sets for a variety of reasons, but to our knowledge it has not seen widespread use as a proxy for the reliability of inferred gene trees despite connections to phenomena such as long-branch attraction (Felsenstein 1978). In the same vein, formal statistical procedures for assessing fit between model and data are generally underutilized. Despite the fact that posterior prediction was introduced to phylogenetics over a decade ago (Bollback 2002) and parametric bootstrapping nearly a decade before that (Goldman 1993a, 1993b), they are still not as widely used as they might be. New test statistics that build on these approaches allow assessment of model fit to focus directly on topology or other parameters of interest (Brown 2014) and, as such, should be of more widespread appeal. However, these new approaches have not yet been broadly tested with empirical data.

Clock-likeness performed better than did topological posterior prediction for the yeast data set, whereas the opposite was true for the amniote UCE data set. The reason for this difference is not immediately obvious, but may reflect differences in the causes of systematic error in the data sets tested or differences in the discriminatory power of the different criteria. Qualitative visual inspection of the distributions of criterion values (Fig. S1 available as the Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.fd3m4) suggests that the criterion that performed best for each data set (clock-likeness for yeast and topological posterior prediction for amniotes) had the most even distribution of values across genes. However, the removal of outliers may weaken this distinction. Alternatively, clock-based filtering may have performed better for the yeast data than the UCE data simply because of the overall higher levels of divergence in the yeast protein-coding genes. Systematic biases caused by unmodeled processes (e.g., heterotachy) may only become apparent above a minimum level of divergence. The difference in performance of posterior prediction across data sets could potentially be explained by soft polytomies present in the yeast reference tree. We employed topological test statistics that evaluate phylogenetic information with respect to the entire tree, but only a subset of splits in the tree are used when comparing to the reference. If posterior prediction is strongly influenced by the effects of model fit on support for splits not present in the reference, which are also the splits that have proven more difficult to definitively resolve, the relationship between PPES and distance to the reference may appear weaker than it actually is. Any remaining paralogs in the yeast data set may also obscure this relationship. Clearly, the behavior of these criteria needs to be characterized across a much wider range of both simulated and empirical data sets, although reference topologies will often not be available to directly compare criterion rankings to the reliability of gene trees.

Rate of evolution was not a very good predictor of gene tree reliability or congruence for either of the exemplar data sets. Surprisingly, comparisons to reference trees produced only very weak positive, or even negative, correlations between rate ranks and those based on distance to references. For amniotes, this correlation was very strongly negative (fast genes were most similar to the reference) and may have been driven by increased stochastic error among more slowly evolving genes, if such genes did not contain a sufficient number of changes to resolve all splits. However, were stochastic error the sole explanation, one might expect a strong negative correlation between rate and intradecile tree-to-tree distances, as well as no correlation between either of the other criteria and distance to the reference. In contrast, the correlation between rate and intradecile ranks is weak (Fig. S5c available as the Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.fd3m4) and the correlation between PPES and reference distance ranks is very strongly positive (Fig. 3b). Rate and PPES also exhibit only a weak negative rank correlation for amniotes (Fig. S11b available as the Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.fd3m4), indicating that these two criteria are characterizing different properties of genes. Because rate itself is a compound function of the state space available at each site (i.e., the set of states with tolerable fitness effects), the proportion of sites-free-to-vary (i.e., those sites with available state spaces greater than 1), and the intrinsic substitution rate, observed variation in rates across genes from each data set may confound easy interpretation. Rate variation in one data set may be driven primarily by differences in mutation rate, whereas it may be driven by complicated and changing patterns of constraints in another. It is noteworthy that the amniote data set was based on UCEs, whose molecular evolutionary constraints and dynamics remain poorly characterized, whereas the yeast data set was based on protein coding sequences whose evolutionary dynamics are better understood.

Splits in the reference trees that exhibited significantly less conflict for preferred genes (more clock-like or lower PPES) were widespread throughout the tree, as were splits that exhibited significantly greater conflict for less preferred genes. However, the modest number of tips in the amniote reference and uncertainty about some parts of the yeast reference (represented as soft polytomies) leads to a small number of internal splits in both cases. Consequently, we are hesitant to draw general conclusions about whether any criterion is likely to improve topological accuracy on splits with particular lengths or occurring at particular positions in the trees.

We have not attempted a full comparison among all possible methods for ranking genes here, but have rather used a comparison between two less utilized approaches (clock-likeness and topological posterior prediction) and one commonly used approach (rate) to highlight the need for further work in this area. Recently, Salichos and Rokas (2013) suggested that commonly used approaches for filtering genes (e.g., alignment certainty, missing data, and rate of evolution) are not

effective in improving phylogenetic congruence across gene trees and that researchers should prefer genes that provide strong support across many splits. However, Salichos and Rokas also noted that this strategy was not effective at increasing gene tree congruence for the most challenging splits. Such splits may be those most affected by systematic biases, as opposed to stochastic error. Contrary to Salichos and Rokas, our results suggest that preferring genes with properties other than the overall amount of support they provide may improve both the reliability and congruence of gene trees, but that the effectiveness of different filtering approaches varies by data set. In agreement with Salichos and Rokas, we found that focusing on small subsets of slowly evolving genes did not increase gene tree reliability or congruence. Betancur-R (2014) highlighted the role of stochastic error in driving incongruence among individual gene trees for slowly evolving genes, and this same effect may explain why filtering by rate of evolution was ineffective in our study. However, filtering by clock-likeness and PPES did improve reliability and congruence across individual gene trees, suggesting that they may minimize systematic error while avoiding the same degree of increased stochastic error inherent to slowly evolving genes. We also note that posterior prediction is a very flexible statistical approach. Alternative test statistics, even others still based on topological information, may have more robust performance across data sets than those we have employed here. Future work will include a broad survey of the relative performance of these statistics. Our posterior predictive results should be considered provisional.

Genes that produce the most topologically unreliable and incongruent phylogenetic estimates may be of great biological interest. Convergence, whether occurring at a small number of positions associated with the function of individual genes (Castoe et al. 2009) or at broader scales due to changes in base composition (Boussau et al. 2008; Nabholz et al. 2011; Betancur-R 2013), frequently misleads attempts to reconstruct phylogenies. However, convergence is also an indicator of repeated adaptive evolution, suggesting that phylogenetically misleading genes may serve as useful starting points for studies of adaptation. The proxies for phylogenetic reliability that we have explored here (clock-likeness and topological PPES) may also identify genes worthy of further scrutiny even when gene tree topologies are accurate. The causes of rate heterogeneity across lineages are of inherent biological interest for a variety of reasons (Lanfear et al. 2010), and patterns of variation in clock-likeness across genes may shed light on the drivers of rate heterogeneity. In addition, topological posterior predictive tests should be able to identify genes that have evolved according to a variety of unexpected evolutionary processes, as long as the patterns generated by these processes are distinct from those consistent with the assumed model of evolution. Examination of posterior predictive outliers may provide biologically inspired avenues for extending models of sequence evolution.

Systematic error (i.e., inconsistency) is particularly concerning in the era of high-throughput sequencing, because one may become increasingly confident in incorrect inferences as more data are gathered. However, such error may be avoided without the loss of much phylogenetic resolution if useful proxies are available for choosing subsets of the data that provide reliable phylogenetic estimates with currently available models. Additionally, much can be learned about the biology of the genome by examining how genes are ranked by such proxies. We join many previous studies (e.g., Jeffroy et al. 2006; Salichos and Rokas 2013; Betancur-R 2014) in suggesting that concatenated or "total evidence" results should be treated with caution, internal conflict across different subsets of the data should be explored and exploited to advance biological understanding, and future work should investigate unexplored properties of genes (or other data subsets) that may be used as proxies for phylogenetic reliability.

Supplementary Material

Supplementary material, including figures and scripts, can be found in the Dryad data repository http://dx.doi.org/10.5061/dryad.fd3m4.

Acknowledgments

Most analyses were conducted with high-performance computational resources provided by Louisiana State University (http://www.hpc.lsu.edu, last accessed June 30, 2015). We would particularly like to acknowledge the support of Jim Lupo from the LSU Center for Computation and Technology. Jaqueline Hess kindly provided the yeast data.

Funding

This research was supported by funds from the LSU College of Science and National Science Foundation award DEB-1355071 to J.M.B.

References

Akaike H. 1974. A new look at the statistical model identification. IEEE Trans. Automat. Contr. 19:716–723.

Benjamini Y., Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B. 57:289–300.

Best D.J., Roberts D.E. 1975. The upper tail probabilities of Spearman's rho. J. R. Stat. Soc. Ser. C. 24:377–379.

Betancur-R.R., Li C., Munroe T.A., Ballesteros J.A., Ortí G. 2013. Addressing gene tree discordance and non-stationarity to resolve a multi-locus phylogeny of the flatfishes (Teleostei: Pleuronectiformes). Syst. Biol. 62:763–785.

Betancur-R.R., Naylor G.J.P., Ortí G. 2014. Conserved genes, sampling error, and phylogenomic inference. Syst. Biol. 63:257–262.

Bogdanowicz D., Giaro K. 2012. Matching split distance for unrooted binary phylogenetic trees. IEEE/ACM Trans. Comput. Biol. Bioinform. 9:150–160.

Bogdanowicz D., Giaro K. 2013. On a matching distance between rooted phylogenetic trees. Int. J. Appl. Math. Comput. Sci. 23: 669–684.
Bogdanowicz D., Giaro K., Wróbel B. 2012. TreeCmp: comparison of trees in polynomial time. Evol. Bioinforma. 8:475–487.
Bollback J.P. 2002. Bayesian model adequacy and choice in phylogenetics. Mol. Biol. Evol. 19:1171–1180.
Boussau B., Blanquart S., Necsulea A., Lartillot N., Gouy M. 2008. Parallel adaptations to high temperatures in the Archaean eon. Nature 456:942–945.
Brinkmann H., van der Giezen M., Zhou Y., Poncelin de Raucourt G., Philippe H. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst. Biol. 54:743–757.
Brown J.M. 2014. Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. Syst. Biol. 63:334–348.
Brown J.M., ElDabaje R. 2009. PuMA: Bayesian analysis of partitioned (and unpartitioned) model adequacy. Bioinformatics 25:537–538.
Brown J.M., Lemmon A.R. 2007. The importance of data partitioning and the utility of Bayes factors in Bayesian phylogenetics. Syst. Biol. 56:643–655.
Castoe T.A., de Koning A.J., Kim H.-M., Gu W., Noonan B.P., Naylor G., Jiang Z.J., Parkinson C.L., Pollock D.D. 2009. Evidence for an ancient adaptive episode of convergent molecular evolution. Proc. Natl. Acad. Sci. USA 106:8986–8991.
Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17:540–552.
Chiari Y., Cahais V., Galtier N., Delsuc F. 2012. Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria). BMC Biol. 10:65.
Crawford N.G., Faircloth B.C., McCormack J.E., Brumfield R.T., Winker K., Glenn T.C. 2012. More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. Biol. Lett. 8:783–786.
Dunn C.W., Hejnol A., Matus D.Q., Pang K., Browne W.E., Smith S.A., Seaver E., Rouse G.W., Obst M., Edgecombe G.D., Sørensen M.V, Haddock S.H.D., Schmidt-Rhaesa A., Okusu A., Kristensen R.M., Wheeler W.C., Martindale M.Q., Giribet G. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745–749.
Edgar R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.
Faircloth B.C., McCormack J.E., Crawford N.G., Harvey M.G., Brumfield R.T., Glenn T.C. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst. Biol. 61:717–726.
Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.
Felsenstein J. 1981. Evolutionary trees from DNA sequences, a maximum likelihood approach. J. Mol. Evol. 17:368–376.
Felsenstein J. 2004. Inferring phylogenies. Sunderland, MA: Sinauer.
Gee H. 2003. Evolution: ending incongruence. Nature 425:782.
Goldman N. 1993a. Statistical tests of models of DNA substitution. J. Mol. Evol. 36:182–198.
Goldman N. 1993b. Simple diagnostic statistical tests of models for DNA substitution. J. Mol. Evol. 37:650–661.
Hess J., Goldman N. 2011. Addressing inter-gene heterogeneity in maximum likelihood phylogenomic analysis: yeasts revisited. PLoS One 6:e22783.
Jeffroy O., Brinkmann H., Delsuc F., Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22:225–231.
Katoh K., Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. Brief. Bioinform. 9:286–298.
Kent W.J. 2002. BLAT - the BLAST-like Alignment Tool. Genome Res. 12:656–664.
Kimura M. 1964. Diffusion models in population genetics. J. Appl. Probab. 1:177–232.
Kimura M. 1968. Evolutionary rate at the molecular level. Nature 217:624–626.
Kumar S., Filipski A.J., Battistuzzi F.U., Kosakovsky Pond S.L., Tamura K. 2012. Statistics and truth in phylogenomics. Mol. Biol. Evol. 29:457–472.
Lanfear R., Welch J.J., Bromham L. 2010. Watching the clock: studying variation in rates of molecular evolution between species. Trends Ecol. Evol. 25:495–503.
Lartillot N., Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21:1095–1109.
Lartillot N., Philippe H. 2008. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 363:1463–1472.
Lemmon A. 2007. MrConverge. Program distributed by the author.
Lin Y., Rajan V., Moret B.M.E. 2012. A metric for phylogenetic trees based on matching. IEEE/ACM Trans. Comput. Biol. Bioinform. 9:1014–1022.
Lopez P., Brinkmann H., Budin K., Laurent J., Moreira D., Guyader L. 2000. Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions. Proc. R. Soc. B Biol. Sci. 267: 1213–1221.
Maddison W.P. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.
Nabholz B., Künstner A., Wang R., Jarvis E., Ellegren H. 2011. Dynamic evolution of base composition: causes and consequences in avian phylogenomics. Mol. Biol. Evol. 28:2197–2210.
Nosenko T., Schreiber F., Adamska M., Adamski M., Eitel M., Hammel J., Maldonado M., Müller W.E.G., Nickel M., Schierwater B., Vacelet J., Wiens M., Wörheide G. 2013. Deep metazoan phylogeny: when different genes tell different stories. Mol. Phylogenet. Evol. 67:223–233.
Nozaki H., Iseki M., Hasegawa M., Misawa K., Nakada T., Sasaki N., Watanabe M. 2007. Phylogeny of primary photosynthetic eukaryotes as deduced from slowly evolving nuclear genes. Mol. Biol. Evol. 24:1592–1595.
Nylander J.A.A. 2004. MrModelTest. Program distributed by the author.
Pamilo P., Nei M. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5:568–583.
Penny D., Foulds L.R., Hendy M.D. 1982. Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. Nature 297:197–200.
Philippe H., Brinkmann H., Lavrov D.V., Littlewood D.T.J., Manuel M., Wörheide G., Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9:e1000602.
Philippe H., Delsuc F., Brinkmann H., Lartillot N. 2005. Phylogenomics. Annu. Rev. Ecol. Evol. Syst. 36:541–562.
Philippe H., Derelle R., Lopez P., Pick K., Borchiellini C., Boury-Esnault N., Vacelet J., Renard E., Houliston E., Quéinnec E., Da Silva C., Wincker P., Le Guyader H., Leys S., Jackson D.J., Schreiber F., Erpenbeck D., Morgenstern B., Wörheide G., Manuel M. 2009. Phylogenomics revives traditional views on deep animal relationships. Curr. Biol. 19:706–712.
Philippe H., Douady C.J. 2003. Horizontal gene transfer and phylogenetics. Curr. Opin. Microbiol. 6:498–505.
Phillips M.J., Delsuc F., Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol. Biol. Evol. 21:1455–1458.
R Core Development Team. 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.
Rambaut A., Grassly N.C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–238.
Robinson D.F., Foulds L.R. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.
Rokas A., Williams B.L., King N., Carroll S.B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.
Ronquist F., Huelsenbeck J.P. 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572–1574.
Ronquist F., Teslenko M., van der Mark P., Ayres D.L., Darling A., Höhna S., Larget B., Liu L., Suchard M.A., Huelsenbeck J.P. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61:539–542.
Salichos L., Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497:327–331.

Schierwater B., Eitel M., Jakob W., Osigus H.-J., Hadrys H., Dellaporta S.L., Kolokotronis S.-O., Desalle R. 2009. Concatenated analysis sheds light on early metazoan evolution and fuels a modern "urmetazoon" hypothesis. PLoS Biol. 7:e20.

Steel M., Penny D. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol. Biol. Evol. 17:839–850.

Sukumaran J., Holder M.T. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:1569–1571.

Swofford D.L. 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sunderland, MA: Sinauer Associates.

Tillier E.R.M., Collins R.A. 1995. Neighbor joining and maximum likelihood with RNA sequences: addressing the interdependence of sites. Mol. Biol. Evol. 12:7–15.

Wapinski I., Pfeffer A., Friedman N., Regev A. 2007. Automatic genome-wide reconstruction of phylogenetic gene trees. Bioinformatics 23:i549–i558.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39:306–314.

Zhong B., Deusch O., Goremykin V.V., Penny D., Biggs P.J., Atherton R.A., Nikiforova S.V., Lockhart P.J. 2011. Systematic error in seed plant phylogenomics. Genome Biol. Evol. 3: 1340–1348.

Zwickl D.J. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence data sets under the maximum likelihood criterion [Ph.D. dissertation]. The University of Texas at Austin.