Syst. Biol. 71(3):676–689, 2022

DOI:10.1093/sysbio/syab075
Advance Access publication September 11, 2021

Excluding Loci With Substitution Saturation Improves Inferences From Phylogenomic Data

David A. Duchêne[1,*], Niklas Mather[2], Cara Van Der Wal[2], and Simon Y.W. Ho[2]

[1]*Centre for Evolutionary Hologenomics, University of Copenhagen, Øster Farimagsgade 5A, 1352 Copenhagen, Denmark; and*
[2]*School of Life and Environmental Sciences, University of Sydney, Sydney, NSW 2006, Australia*
[*]*Correspondence to be sent to Centre for Evolutionary Hologenomics, University of Copenhagen, Øster Farimagsgade 5A, 1352 Copenhagen, Denmark; E-mail: david.duchene@sund.ku.dk.*

Received 19 April 2020; reviews returned 00 Month 2021; accepted 7 September 2021
Associate Editor: James Rosindell

Abstract.—The historical signal in nucleotide sequences becomes eroded over time by substitutions occurring repeatedly at the same sites. This phenomenon, known as substitution saturation, is recognized as one of the primary obstacles to deep-time phylogenetic inference using genome-scale data sets. We present a new test of substitution saturation and demonstrate its performance in simulated and empirical data. For some of the 36 empirical phylogenomic data sets that we examined, we detect substitution saturation in around 50% of loci. We found that saturation tends to be flagged as problematic in loci with highly discordant phylogenetic signals across sites. Within each data set, the loci with smaller numbers of informative sites are more likely to be flagged as containing problematic levels of saturation. The entropy saturation test proposed here is sensitive to high evolutionary rates relative to the evolutionary timeframe, while also being sensitive to several factors known to mislead phylogenetic inference, including short internal branches relative to external branches, short nucleotide sequences, and tree imbalance. Our study demonstrates that excluding loci with substitution saturation can be an effective means of mitigating the negative impact of multiple substitutions on phylogenetic inferences. [Phylogenetic model performance; phylogenomics; substitution model; substitution saturation; test statistics.]

One of the key steps in phylogenomics is to identify a suitable set of loci for reconstructing the evolutionary history of a group of organisms. The inference of phylogenetic trees from nucleotide sequences can be misled by a number of factors. For example, the sequences might contain too little information if they have evolved too slowly (Yang 1998; Klopfstein et al. 2017). On the other hand, high evolutionary rates will increase the probability of multiple substitutions occurring at the same nucleotide sites, leading to a phenomenon known as substitution saturation (Brown et al. 1982; Mindell and Honeycutt 1990; Philippe and Forterre 1999; Philippe et al. 2011). Even when the best-fitting model of nucleotide substitution is used, saturation can cause the phylogenetic method to produce inaccurate estimates of the tree topology and branch lengths. Therefore, an important step in experimental design for phylogenomic analysis is to identify the loci with excessive evolutionary rates that might mislead phylogenetic inference (Philippe et al. 2011).

One popular approach for exploring phylogenetic informativeness is to identify the loci that have experienced substitution saturation (Philippe and Forterre 1999). A widely used approach for this purpose is to compare the sites in a sequence alignment with those expected under conditions of complete saturation. Following information theory, the test takes the pattern of nucleotide frequencies at a fully saturated site to follow a multinomial distribution with maximum entropy (Xia et al. 2003). The critical values for this test are the entropy values above which the estimates of the tree topology and branch lengths are likely to be inaccurate.

In an extensive simulation study, Xia et al. (2003) described the behavior of the entropy test across a range of phylogenetic conditions of overall evolutionary rate, amounts of data, and tree imbalance. However, there is a wide range of factors that can interact to mislead phylogenetic inference, such as substitution model underparameterization (Revell et al. 2005; Sullivan and Joyce 2005) or the presence of long terminal branches (Klopfstein et al. 2017; Dornburg et al. 2019). Some of these factors are rapidly gaining recognition in phylogenomics research, in which identifying sources of bias can be crucial for obtaining a reliable estimate of the phylogeny (Reddy et al. 2017; Mai and Mirarab 2018). Therefore, understanding the sensitivity of tests of saturation to a broad range of phenomena in empirical data can improve practice in phylogenomics.

Data sets in phylogenomics are typically composed of many alignments of non-recombining regions of the genome (loci). The phylogenetic information signal in each locus is often summarized using its "gene tree." Tests of substitution saturation can be used to select loci for phylogenetic analysis (e.g., Han and Ro 2015; Dávalos and Perkins 2008; Liu et al. 2014). This form of data filtering ultimately aims to maximize the signal of the true phylogenetic relationships in the data, also known as the historical signal. There has been growing interest in data-filtering methods for phylogenomics (Molloy and Warnow 2018; Richards et al. 2018; Bravo et al. 2019). However, the effectiveness of these methods, such as using tests of saturation, remains to be explored in depth.

In this study, we describe the performance of a common test of saturation, and evaluate the impact of saturation in a broad range of empirical phylogenomic data sets. We also describe a novel approach to examining substitution saturation that focuses exclusively on phylogenetically informative sites. This approach greatly ameliorates the negative influence of slowly evolving sites on the measurement of overall base composition, which is central to the performance of the entropy-based tests of saturation. Using two simulation studies, we first aim to identify the characteristics of sequence alignments to which the tests are sensitive, including a high rate of substitution. In our second simulation study, we explore a broad continuum of evolutionary scenarios to examine the power of the tests to identify loci with amounts of substitution saturation that are likely to mislead estimates of phylogeny and branch lengths. We then evaluate the degree of substitution saturation in 36 phylogenomic data sets, and investigate the link between saturated loci and amount of discordance in phylogenetic signal across sites. Our results suggest that saturation can affect large portions of phylogenomic data sets, and that testing for saturation is an effective approach to identify loci with poor historical signal in phylogenomic studies.

Materials and Methods

Test of Expected Entropy

The degree of substitution saturation in a nucleotide sequence alignment can be described using a measure of entropy (Xia et al. 2003). The entropy of a distribution is defined as the average information content in a given sample. The information content measures the level of surprise expected when we encounter a particular outcome. Systems with high entropy are more unpredictable and disorderly. In the context of phylogenetics, entropy is highest at full saturation, when noise has overridden the historical signal. In neutrally evolving sequences under full saturation, every site might be expected to have nucleotide frequencies equal to those of the whole alignment. The nucleotide that occurs in each sequence at a site is assumed to be independent of the nucleotides in other sequences, so that the vector of nucleotides at each site can be modeled as a single draw from a multinomial distribution, where the underlying probabilities are equal to the nucleotide frequencies. The entropy of a multinomial sample with n observations, k categories, and a vector p of probabilities for each of the categories is:

$$H(X) = -\log(n!) - n\sum_{i=0}^{k} p_i \log(p_i) + \sum_{i=0}^{k}\sum_{x_i=0}^{n}\binom{n}{x_i} p_i^{x_i}(1-p_i)^{n-x_i}\log(x_i!) \qquad (1)$$

Therefore, the expected entropy of a sequence alignment under full saturation can be calculated using the number of taxa (n) and the overall nucleotide frequencies (p). Meanwhile, the information content at an observed site is based on the counts of nucleotides at that site:

$$I(X|X=x) = I(x) = -\log_2(P(X=x)), \qquad (2)$$

where $P(X=x)$ is the density of a sample from the multinomial distribution, given that the probabilities are the overall nucleotide frequencies in the sequence alignment. Since the entropy is the expected value of the information content, we can calculate the entropy of our sample by taking the average of the information at each site. A test of saturation can then be performed by taking the estimates of entropy across alignment sites and comparing this with the sample entropy expected under full saturation.

Substitution saturation can have negative impacts on phylogenetic inference even before the historical signal is completely eroded (Ho and Jermiin 2004), so a classic test of the significance of the distance (t) between the distribution of observed site information against the value of maximal entropy will provide only limited information about phylogenetic inferences. Instead, a critical value for the t-statistic test can be derived from simulations (Xia et al. 2003). The critical value t_{crit} can then be used for testing the hypothesis that the empirical value of t_{obs} is far from maximal entropy, a situation in which we would expect a negligible impact of substitution saturation on phylogenetic inference. Our formulation of this test is slightly different from that of Xia et al. (2003), who tested whether t_{obs} is significantly smaller than t_{crit}. We propose testing whether t_{obs} is significantly smaller than full entropy, using t_{crit} as the critical value of the test. Our test considers the variance in the information content across sites more explicitly, but the two tests can be expected to produce similar results.

Existing implementations of the test of saturation take the expected entropy to depend on the base frequencies of the sequence alignment. However, this test can have problematic behavior in several scenarios, such as when sequences are subject to selective constrains or when sites vary in evolutionary rate. Such variation in rates is common in empirical data sets: among codon positions in protein-coding sequences or across sites in ultraconserved elements (UCEs) and their flanking regions. Therefore, we assessed the performance of the test based exclusively on phylogenetically informative sites (i.e., parsimony-informative sites with at least two nucleotides, where at least two of these occur in at least two sequences). This avoids the worst effects of constant sites on the test in data matrices with small numbers of taxa, small numbers of sites, or both. The focus of the test on phylogenetically informative sites will have a diminishing effect on the test as the size of the data matrix increases, and is expected to have a null effect under infinite sites.

The guidelines require revision so that the test can be implemented more effectively for empirical data. Very small sequence alignments can pose challenges for the reliability of the test, due to the disproportionate effect of small numbers of outlier sites. Similarly, alignments with large numbers of slow-evolving sites and an overrepresentation of a particular nucleotide type will be highly sensitive to single outlier sites. Therefore, a test of saturation using the entropy t-statistic is unlikely to perform well on small sequence alignments or in which nucleotide frequencies are highly uneven.

The Entropy t-Statistic and Phylogenetic Inference

We used a simulation study to examine the circumstances under which substitution saturation is a significant problem that can be addressed in empirical data, relative to other causes of misleading inferences. We simulated the evolution of nucleotide sequences along trees with variable numbers of taxa (8, 32, 128, and 512) to generate data sets with three different sequence lengths (250, 500, and 1500 nucleotides). Starting with trees that were fully symmetric and in which all branches had equal length, we varied several conditions that can affect the performance of phylogenetic methods, including: the number of substitutions along each branch of the tree (0.05, 0.25, 0.45, and 0.65); tree imbalance (fully balanced tree versus fully imbalanced tree); ratio of the sum of internal to the sum of external branch lengths, or stemminess (0.1, 0.5, and 0.9; Fiala and Sokal 1985); and the substitution model that generated the data (the matched Jukes–Cantor model, JC; or the more complex GTR+Γ model). Simulations under GTR+Γ focused on examining the impact of model underparameterization, and were made with transition parameters drawn from a Dirichlet distribution (all values of $\alpha = 5$) and gamma-distributed rates across sites with $\alpha = 1$. In addition, we simulated sequence evolution under the conditions above but including a proportion of constant sites (0, 0.25, 0.5, 0.75), as is common in several empirical data types. Our simulations included 100 sequence alignments under each combination of scenarios and under either the JC or the GTR+Γ substitution model. We performed phylogenetic analyses of each data set using the JC substitution model and maximum-likelihood optimization in IQ-TREE (Nguyen et al. 2015).

Phylogenetic accuracy in each scenario was calculated as the unweighted and normalized Robinson–Foulds distance (Robinson and Foulds 1981; Penny and Hendy 1985) between the inferred tree and the "true" tree used for simulation in each analysis. We also made a comparable calculation with branch lengths, by taking the difference between the estimated and true tree length, and dividing this difference by the true tree length. The outcome is the proportion error in estimated tree length. In addition, we recorded the entropy *t*-statistics as calculated on all alignment sites and on informative sites only.

We tested the hypothesis that each factor that we varied in our simulations had an impact on phylogenetic inference and on the two entropy *t*-statistics. We tested four linear regression models in which the response variables were, in turn, the topological distance between the estimated and true trees, the difference in tree length between the estimated and true trees, and each of the two t_{obs} values of the saturation test (calculated on all sites or only on informative sites). The explanatory variables were the six fixed factors that we varied in our simulations (number of taxa, alignment length, tree length, degree of tree imbalance, stemminess, and substitution model used for simulation). In addition, we considered the total number of variable sites in the alignment, and all of the two-way interactions between the main effects. The resulting *P*-values were corrected for multiple comparisons using false discovery rates.

Diagnostic Ability of the Entropy t-Statistic

In a second simulation study, we characterized the performance of the two entropy *t*-statistics for identifying cases in which substitution saturation is misleading phylogenetic inference. We explored a similar parameter space as in our first simulation study, but treated several of the factors as continuous. In each simulation, we sampled values from uniform distributions for mean branch length U(0.01, 0.65), stemminess U(0.1, 0.9), and sequence length U(250, 2000). In each simulation, we also randomly sampled the number of taxa (8, 32, 128, and 512), whether the tree was imbalanced or balanced, a portion of the sites to remain constant U(0, 0.8), and the model of nucleotide substitution used for sequence evolution (JC or GTR+Γ). We repeated 10^5 times the sampling of parameters, simulation of sequence evolution, phylogenetic inference in IQ-TREE under a JC model, and the calculation of the two *t*-statistics of entropy (all sites versus informative sites only). This analysis allowed us to explore the performance of the test statistics along a gradient of scenarios and to identify appropriate critical values for interpreting the best-performing test statistic.

We quantified the diagnostic ability of the two entropy *t*-statistics across the simulation conditions by using receiver operating characteristic (ROC) curves to show the relationship between rates of true positives and false positives. Across a traverse of the values of the test statistic calculated in simulations, we considered positives to be cases in which the topology of the inferred phylogeny was different from that of the true tree. For each ROC curve, we chose the critical values as those that maximized the difference between the numbers of true positives and false positives, such that they provide the greatest power for discriminating positives from negatives. The code used to perform simulations is freely available online (github.com/duchene/entropy_saturation_test).

We also explored the performance of the entropy *t*-statistics in identifying cases in which substitution saturation is likely to have misled estimates of branch lengths. Branch lengths are continuous variables, so an arbitrary critical value has to be used to determine the target positives for the test. We determined a positive as being cases in which estimated tree length was at least 50% greater or smaller than the true tree length.

Based on the results of our simulations, we propose critical values for assessing saturation using the entropy test statistic calculated on phylogenetically informative sites. Past research has indicated that test statistics often cannot be interpreted using a single, universally applied critical value, because the power of any given test can vary with sample size and other features of the data (e.g., Xia et al. 2003; Duchêne et al. 2017). Following previous work, we propose critical values that depend on the number of taxa and on sequence length. Specifically, we performed a multiple regression of the critical value chosen under each simulation condition on the square root of the corresponding number of taxa and sequence length as explanatory variables. Regression models can then be used to predict critical values across any number of taxa and sequence length. The saturation test is implemented in the free software PhyloMAd (Duchêne et al. 2018b; github.com/duchene/phylomad) and reports the test results, a diagnosis based on the new critical values, and estimates of expected false-positive and true-positive rates.

Signal of Substitution Saturation in Phylogenomic Data

The prevalence and impact of substitution saturation were examined in a range of phylogenomic data sets from the literature (Table 1). We obtained 36 data sets that varied widely in taxonomic range, size, and sequence type. The data sets spanned multiple orders of magnitude in their number of taxa and number of genomic regions included. They also included a range of data types that we broadly describe as exon, intron, ultraconserved element, or anchored-enriched region data. While these data types have some overlap, they are different in the data targeted by researchers (e.g., coding gene regions versus the informative regions flanking UCEs), which potentially allows useful distinctions to be made.

Our proposed test of substitution saturation was performed separately for each locus in each data set, where “locus” is defined as in the studies from which we obtained the data. Our analyses focused on the test on informative sites only, given that a test on complete alignments has inferior performance under most conditions (see *Results*). We only explore loci with a minimum of 30 phylogenetically informative sites, and in which none of the nucleotide frequencies is greater than 0.5. These bounds ameliorate the severe impact that outlier sites can have on the estimates of the *t*-statistic. For each data set, we compared loci with high versus low risk of saturation in terms of their total numbers of sites, numbers of informative sites, GC content, and phylogenetic inferences. Phylogenetic analyses were performed for each locus using maximum likelihood in IQ-TREE (Nguyen et al. 2015) with the best-fitting substitution model from the GTR+Γ family. As a metric of branch support, we calculated the approximate likelihood-ratio test (aLRT), which assesses the agreement across sites regarding the maximum-likelihood resolution (Guindon et al. 2010). From the inferences for each locus, we extracted the mean branch support across branches, mean estimated branch lengths, and stemminess (Fiala and Sokal 1985).

Results

The Entropy t-Statistic and Phylogenetic Inference

Our simulations show that substitution saturation, simulated as long tree lengths, is one of the primary factors misleading phylogenetic inference (Fig. 1). However, several other factors are also important obstacles to phylogenetic inference, including tree imbalance, large numbers of taxa, low relative lengths of internal to external branches (stemminess), and high complexity in the true substitution process compared with the model used for inference (model underparameterization; Fig. 1).

Our regression models highlight the widespread interactions between factors in misleading phylogenetic inferences. For example, substitution saturation is highly misleading to topology inference when tree imbalance is high and trees have large numbers of taxa (Fig. 1a,b). Accuracy in the inferred tree topology was best explained by the two-way interactions between number of taxa and phylogenetic imbalance of the true tree (*t*-value $=218.971$; Supplementary Table S1 available on on Zenodo at https://doi.org/10.5281/zenodo.5131558), that between phylogenetic imbalance and stemminess of the true tree (*t*-value $=111.592$), between the true tree length and the substitution model used for simulation (Fig. 1a; *t*-value $=-104.186$), and between the true tree length and stemminess of the true tree (*t*-value $=-86.007$).

Several of the primary factors explaining error in branch-length estimates were the same as those explaining error in the tree topology, including the interaction between true tree length and the substitution model used for simulation (Fig. 1; *t*-value $=-191.171$), and the interaction between tree imbalance and the stemminess of the true tree (*t*-value $=110.811$). In addition, a strong predictor of error in branch-length estimates was the interaction between the substitution model used and the proportion of invariable sites (*t*-value $=215.571$). All *P*-values were small (<0.001) and remained qualitatively identical when adjusted using false-discovery rates.

The entropy *t*-statistics were sensitive to factors that were similar, but not identical, to those that best explained phylogenetic error. The coefficients of regressions explaining topological error and the entropy statistics were similar (Fig. 2a,b). Meanwhile, coefficients of regression explaining error in tree-length estimate were not associated with those explaining the entropy statistics (Fig. 2c,d). Therefore, the saturation test presented here is primarily suited to assessing misleading estimates of tree topology rather than branch lengths.

The *t*-statistic applied to phylogenetically informative sites has a greater sensitivity to factors that also affect phylogenetic inference compared with the statistic on all sites. The statistic calculated on all sites was best explained by the interaction between true tree length and stemminess of the true tree (*t*-value $=144.484$), followed by the interaction between true tree length and phylogenetic imbalance (*t*-value $=114.893$). However,

TABLE 1. Phylogenomic data sets tested for substitution saturation in this study

Taxon	Number of loci[a]	Number of taxa per locus	Data type/ genomic region[b]	Mean run time of saturation test per locus (s)	Source reference
Stinging wasps (Aculeata)	807	140–183	UCE	0.273	Branstetter et al. 2017
Bilaterian metazoans (Bilateria)	424	50–75	Exon	0.076	Cannon et al. 2016
Laurasiatherian mammals (Laurasiatheria)	10,258	8–23	Intron	0.096	Chen et al. 2017 (A)
Laurasiatherian mammals (Laurasiatheria)	3637	5–23	Intron	0.118	Chen et al. 2017 (B)
Amniote vertebrates (Amniota)	1145	10	UCE	0.035	Crawford et al. 2012
Marsupial mammals (Marsupialia)	1494	38–45	Exon	0.098	Duchêne et al. 2018a
Butterflies (Papilionoidea)	350	144–205	Exon	1.086	Espeland et al. 2018
Ray-finned fishes (Actinopterygii)	489	5–27	UCE	0.040	Faircloth et al. 2013
North American tarantulas (*Aphonopelma*)	581	63–83	Anchored	0.179	Hamilton et al. 2016 (A)
Spiders (Araneae)	326	22–34	Anchored	0.107	Hamilton et al. 2016 (B)
North American mygalomorph spiders (Euctenizidae)	403	18–25	Anchored	0.085	Hamilton et al. 2016 (C)
Ray-finned fishes (Actinopterygii)	1101	105–298	Exon	1.095	Hughes et al. 2018
Cichlid fishes (Cichlidae)	533	57–149	Anchored	0.927	Irisarri et al. 2018
Birds (Aves)	8293	42–52	Exon	0.154	Jarvis et al. 2014 (A)
Birds (Aves)	8287	42–52	Exon	0.117	Jarvis et al. 2014 (B)
Birds (Aves)	2515	39–52	Intron	0.540	Jarvis et al. 2014 (C)
Gobioid fishes (Actinopterygii: Gobioidei)	570	43	Exon	0.133	Kuang et al. 2018
Iguanas (Phrynosomatidae)	580	4–11	UCE	0.052	Leaché et al. 2015
Flowering plants (Angiosperms)	370	29–35	Anchored	0.089	Léveillé-Bourret et al. 2018
Mosses (Bryophyta)	105	68–146	Exon	0.290	Liu et al. 2019
Birds (Neoaves)	1539	17–33	UCE	0.041	McCormack et al. 2013
Songbirds (Passeri)	515	106	UCE	0.191	Moyle et al. 2016
Acorn ants (*Temnothorax*)	2091	44–50	UCE	0.124	Prebus 2017
Birds (Aves)	259	164–200	Anchored	0.750	Prum et al. 2015
Snakes (*Storeria*)	322	70–90	Anchored	0.389	Pyron et al. 2016
Seed plants (Gymnosperms)	1308	38	Exon	0.089	Ran et al. 2018 (A)
Seed plants (Gymnosperms)	1308	38	Exon	0.082	Ran et al. 2018 (B)
Seed plants (Gymnosperms)	1308	38	Exon	0.093	Ran et al. 2018 (C)
Harvestmen spiders (Ischiropsalidoidea)	672	5	Exon	0.037	Richart et al. 2016 (A)
Harvestmen spiders (Ischiropsalidoidea)	653	5	Exon	0.034	Richart et al. 2016 (B)
Harvestmen spiders (Ischiropsalidoidea)	672	5	Exon	0.034	Richart et al. 2016 (C)
Ferns (Monilophyta)	2385	52–73	Exon	0.118	Shen et al. 2018
Squamate reptiles (Squamata)	4175	18–34	UCE	0.054	Streicher and Wiens 2017
Squamate reptiles (Squamata)	44	98–167	Exon	0.810	Wiens et al. 2012
Decapod crustaceans (Decapoda)	105	57–94	Exon	0.849	Xia et al. 2003
Squamate reptiles (Squamata)	52	98–2378	Anchored	1.086	Zheng and Wiens 2016

Note: Data sets were examined as done in previous studies, such that each locus was analyzed independently.

[a] For some studies, only a subset of the data were examined because of numerical problems (possibly caused by missing data) or file-format difficulties (such as those caused by unusual characters).

[b] "Exon" and "Intron" data sets refers to gene regions, either protein-coding or intronic, respectively; "UCE" refers to ultraconserved elements (Faircloth et al. 2012); "Anchored" refers to anchored-enriched regions (Lemmon et al. 2012).

the test statistic is weakly correlated with an underparameterized substitution model and extreme cases of tree imbalance (Supplementary Table S1 available on Zenodo). In contrast, the statistic on phylogenetically informative sites was reasonably well explained by the interaction between the true tree length and the substitution model used for simulation (t-value $=124.593$), and between imbalance and stemminess (t-value $=72.962$).

Diagnostic Ability of the Entropy t-Statistic

The entropy t-statistic test focusing on informative sites consistently made a clear separation between true positives and false positives across a traverse of values of the test statistic, suggesting strong power to discriminate between accurate and misleading inferences of tree topology (Fig. 3). In contrast, the test using all alignment sites performed poorly in distinguishing true positives

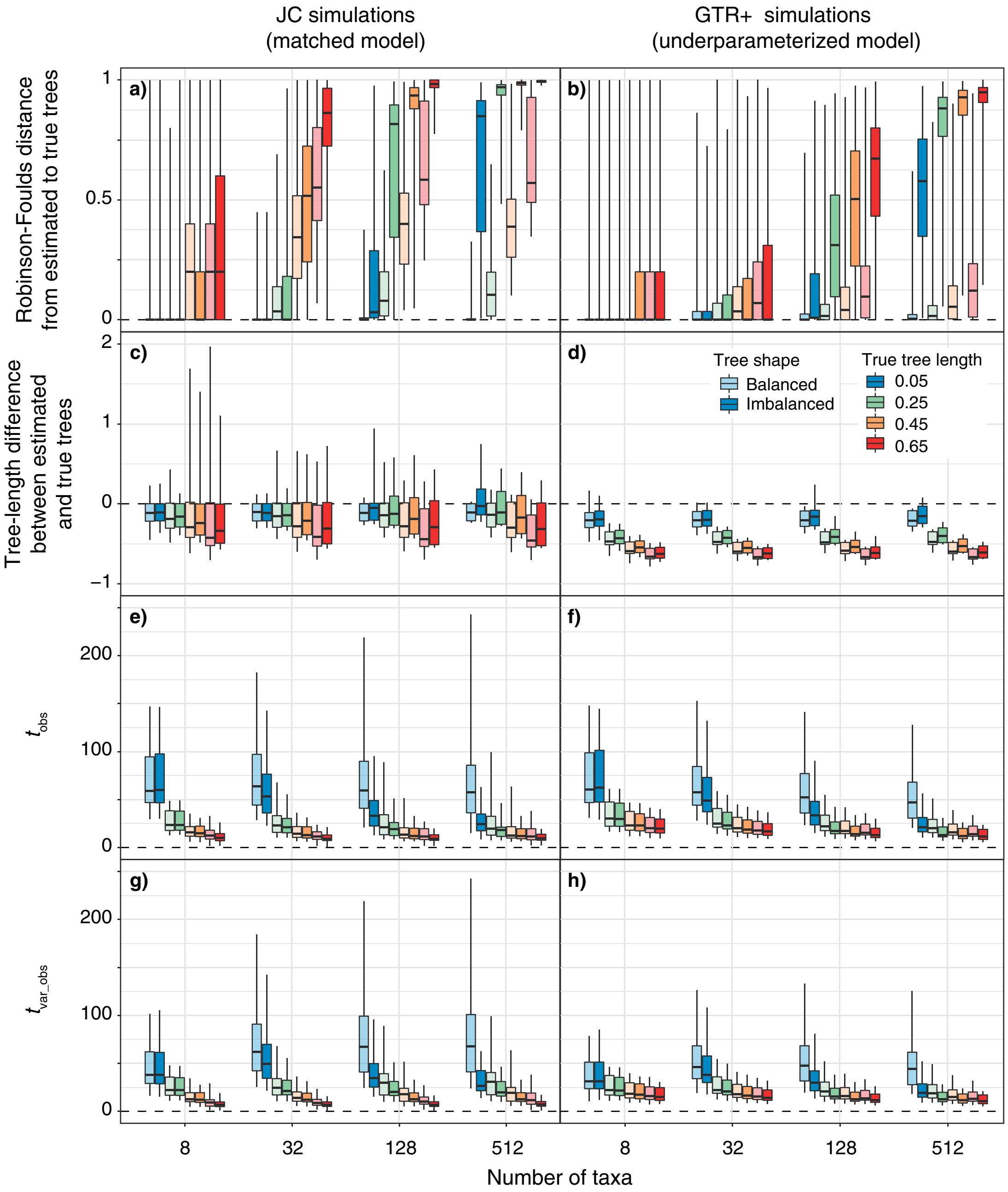

FIGURE 1. The effect of the most influential factors in simulations on the performance of phylogenetic inference and the entropy t-statistic. The JC model of nucleotide substitution was used for all analyses, such that the model is matched in simulations under the JC model (a, c, e, and g), and underparameterized in simulations under the GTR+Γ model (b, d, f, and h). The performance of phylogenetic inference was measured as (a, b) the unweighted Robinson–Foulds distance between the estimated and true trees, and (c, d) the difference in tree length between estimated and true trees, calculated as ((estimated tree length – true tree length) / true tree length). Performance is compared with the entropy t-statistic as calculated on all alignment sites (c, d; t_{obs}) and on phylogenetically informative sites only (c, d; $t_{\text{var\_obs}}$). Boxplot whiskers cover the full range of values in each scenario.

from false positives and was virtually ineffective when the substitution model was underparameterized. Under a matched substitution model, the smallest number of taxa explored (8) led to the poorest test performance, primarily caused by a minority of true negatives having values that resemble most of the true positives (i.e., some accurate inferences having a signal that resembles high entropy; Fig. 3). Substitution model underparameterization primarily affected test performance in analyses with large numbers of taxa. Long sequences led to a small drop in test performance in analyses with large numbers of taxa (128 and 512 taxa).

The number of taxa and sequence length are suitable predictors of the best thresholds as taken from the

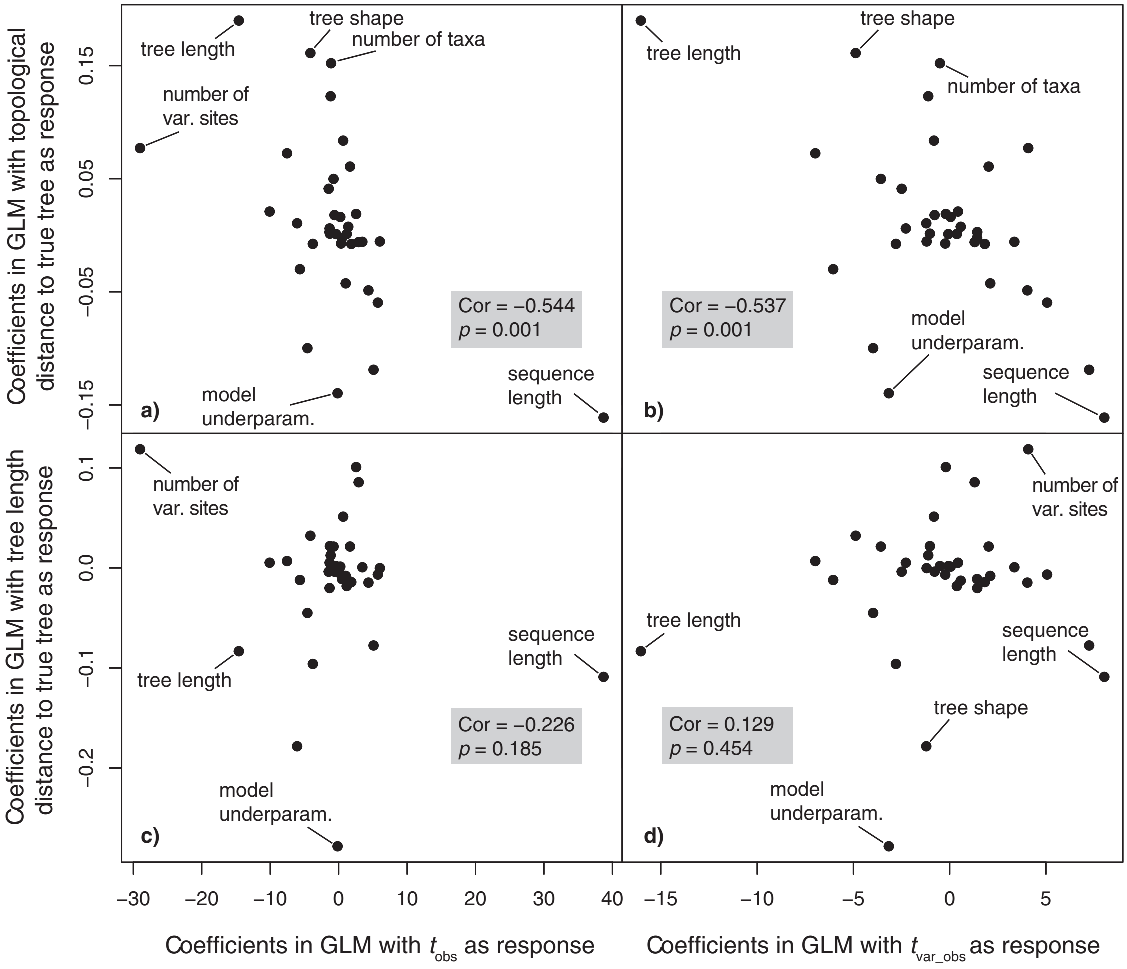

FIGURE 2. Coefficients of regression analyses of multiple variables across simulations explaining topological distance to the true tree and tree length distance (y-axis) and the entropy t-statistics (x-axis). A strong negative correlation between regression coefficients explaining topological distance and entropy t-statistics (a, b) indicates that the statistics have some power to predict misleading inferences of tree topology. Meanwhile, there is no association in regression coefficients explaining tree-length distance to the true tree and the entropy statistics (c, d).

data in ROC curves (the value that maximizes the difference between numbers of true positives and false positives). Specifically, according to multiple regression, the values of t_{crit} of the test on informative sites are well explained by the square root of numbers of taxa and sequence length (adjusted $r^2 = 0.9$; Fig. 4). Nonetheless, there will be excessive uncertainty around the predicted threshold values when the number of taxa is ≫512 and sequence length is ≫1600. Another observation is the high values of t_{crit} at intermediate numbers of taxa. At these tree sizes, estimates of base composition will be the most influenced by sites that are slow-evolving yet are parsimony-informative. For this reason, base composition is the least representative of maximum entropy at intermediate numbers of sites. This is reflected in the high variance in estimates of t_{crit} and higher predicted values at those intermediate numbers of taxa (Fig. 4).

The test on informative sites also had a reasonable ability to identify inaccurate inferences of branch lengths (Supplementary Fig. S1 available on Zenodo), while the test on all sites had poor performance throughout scenarios. The test on informative sites was also virtually unaffected by model underparameterization when including small numbers of taxa. These results are consistent with our simulations showing that the primary drivers of misleading branch-length estimates are large tree lengths, high stemminess, and excessive tree imbalance, rather than the accuracy of the substitution model among those examined.

Signal of Substitution Saturation in Phylogenomic Data

Our implementation of the entropy tests of saturation on empirical data sets took an average of 0.141 s per locus. Saturation was flagged as being highly prevalent in more than a quarter of the data sets examined (10 of 36), affecting >5% of loci in these data sets and >50% in one case. In the remaining data sets, saturation was rarely flagged, or not at all (26 of 36). Given this striking distinction among data sets, our

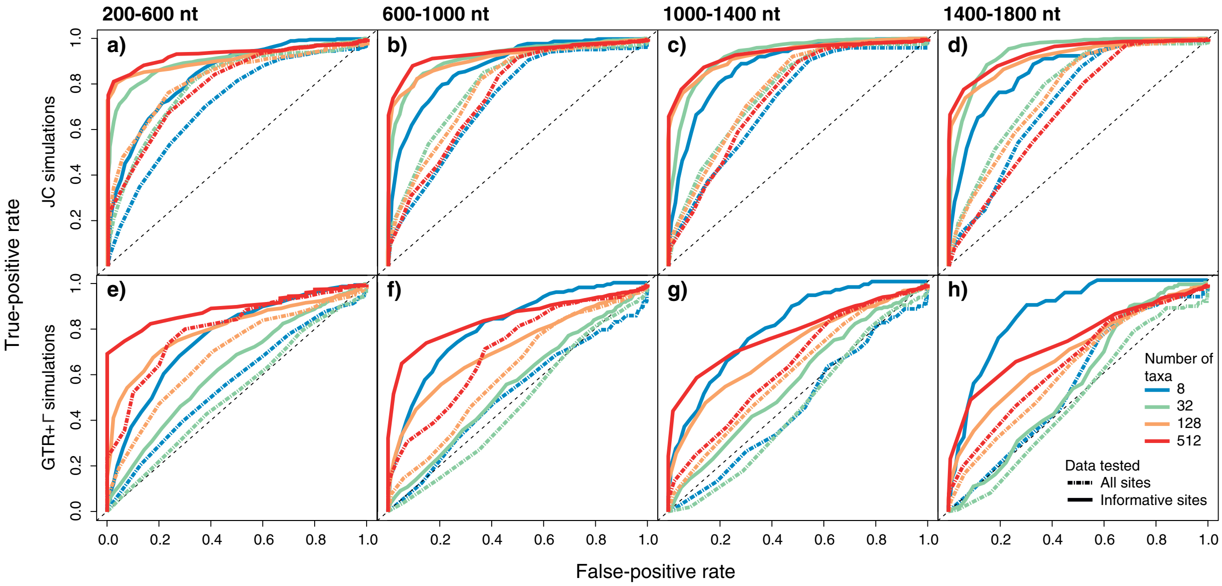

FIGURE 3. Receiver operating characteristic (ROC) curves showing the power of the entropy t-statistic test to classify phylogenetic topological inferences that are accurate versus inaccurate. Panel columns separate the results across four categories of alignment length, each of width 400 nt. Panel rows separate the results of analyses where the model used for simulation matched that used for inference (a–d), versus scenarios with model underparameterization (e–h). The curves show the proportion of true positives and false positives as discriminated across the range of values of the entropy t-statistic. Each line shows the results from a set of approximately 3100 simulations based on random combinations of substitution rates, tree balance/imbalance, and tree stemminess.

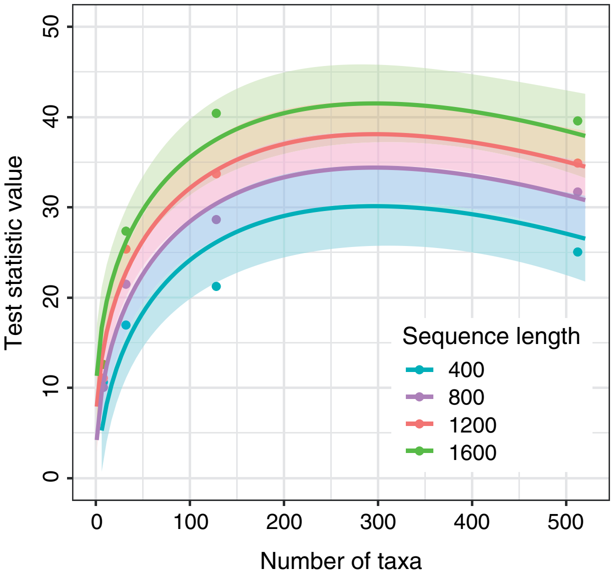

FIGURE 4. Predicted values of t_{crit} using a multiple regression in which the explanatory variables are the square root of sequence length and number of taxa. The predictions are made for a test using exclusively phylogenetically informative sites. Each critical value is chosen to maximize the difference between numbers of true positives and false positives (taken from the data shown in Fig. 3). Lines show the predicted values of the tcrit across different numbers of taxa, separated in color by each of the values of sequence length considered. Shading indicates the uncertainty around the predicted values of tcrit across numbers of taxa and sequence lengths.

analyses show that, when prevalent, saturation can have a large impact on phylogenomic analyses. Branch support (aLRT) was on average lower in saturated than unsaturated loci in 70% of the data sets that had any flagged saturation (Fig. 5a). Unsaturated loci also had a number of informative sites that was high relative to other loci in their corresponding study (Fig. 5b). This means that rather than having an absolute small number of informative sites, it was loci with relatively small numbers of informative sites that tended to be flagged for saturation. Saturated and unsaturated loci led to similar estimates of branch lengths and stemminess, and had comparable GC contents (Supplementary Fig. S2 available on Zenodo).

Our analyses revealed that the impact of saturation on discordance across sites in gene trees can be substantial. Loci flagged for saturation yielded gene trees that had mean aLRT branch supports that were 4.8% lower on average than the trees inferred from unsaturated loci, being 25% poorer in one data set (Fig. 6). Nonetheless, the portion of saturated loci was not associated with the loss in branch support in gene trees from saturated loci (Fig. 6b). The data sets that benefited the least from distinguishing between saturated and unsaturated loci tended to be UCE or anchored-enriched data. These data sets also had a slight tendency to have fewer informative sites per locus and smaller numbers of taxa (Fig. 6c,d).

Data sets that targeted exons and introns had higher proportions of saturated loci than those that targeted UCEs and anchored-enriched regions (Fig. 7a). Data sets from exons/introns also tended to have greater numbers of informative sites and to yield longer gene trees with greater aLRT branch supports than data sets from UCEs and anchored-enriched regions (Fig. 7b,c). In addition, data sets comprising exon/introns led to

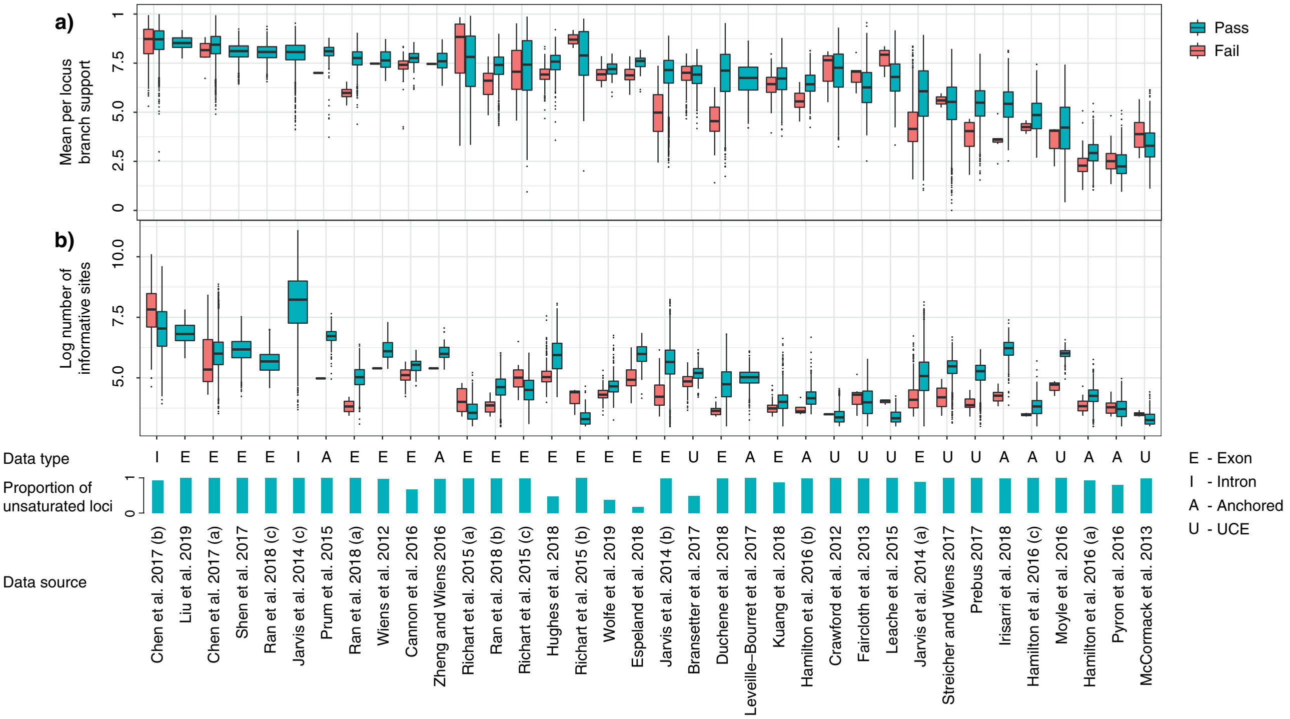

FIGURE 5. Characteristics of loci with and without substitution saturation from 36 phylogenomic data sets: (a) mean branch support in inferred gene trees, and (b) the number of informative sites across loci. Data sets are ordered from left to right by the mean branch support in trees inferred from unsaturated loci. The *x*-axis shows for each study the data type, the proportion of loci found to have no substitution saturation, and the source publication.

gene trees with shorter internal relative to external branch lengths (lower stemminess; Fig. 7e). The longer internal branches found in gene trees inferred from UCEs or anchored-enriched regions suggest that low branch supports are likely associated with the lower number of sites and overall shorter trees, rather than due to misleading inferences such as those arising from substitution saturation (Fig. 5).

DISCUSSION

Our study has shown that substitution saturation can be highly misleading in phylogenetics, even when compared with a range of other factors that can also mislead inference. Saturation was flagged as occurring in large numbers of loci in nearly one-third of a broad range of empirical data sets examined, and tended to be flagged in loci with highly discordant signals across sites (Supplementary Fig. S3 available on Zenodo). Interestingly, loci with relatively few informative sites in each data set were frequently flagged as being saturated, such that they should be scrutinized in empirical data. One solution is to use the entropy *t*-statistic calculated on phylogenetically informative sites to rapidly identify levels of substitution saturation that are likely to mislead phylogenetic inference. Similar tests for identifying historical signal in sequence alignments vary considerably in their effectiveness (e.g., Strimmer and Von Haeseler 1997; Goldman 1998; Townsend 2007; Susko and Roger 2012; Townsend et al. 2012; Klopfstein et al. 2017). For this reason, we have focused on a description of the rates of true positives versus false positives of entropy tests for substitution saturation and a fast implementation. Our results indicate that saturation is a relatively common and problematic phenomenon in phylogenomics; accordingly, the entropy test of saturation as described here offers a useful complement to existing diagnostics of data quality and model performance.

Our analyses of empirical data sets suggest that different data types have comparable portions of saturated loci. Rather than depending on data type, saturation is typically flagged as a problem for loci that have small numbers of fast-evolving sites. This finding lends support to the hypothesis that variable sites in highly conserved genomic regions have more saturation than those in highly variable genome regions (Philippe et al. 1996). Despite the similar results across data types, we also observed only limited improvement in gene-tree branch supports when excluding saturated loci in UCE data sets (Supplementary Fig. S3 available on Zenodo). This might be due to a more careful choice of markers and data filtering before analyses of UCEs, which is also reflected in the small portions of loci rejected.

Strikingly, our test identified large proportions of loci as being saturated in multiple data sets, in some cases identifying saturation in more than half of the data. This suggests that assessing substitution saturation can have a dramatic effect in reducing noise in phylogenomic

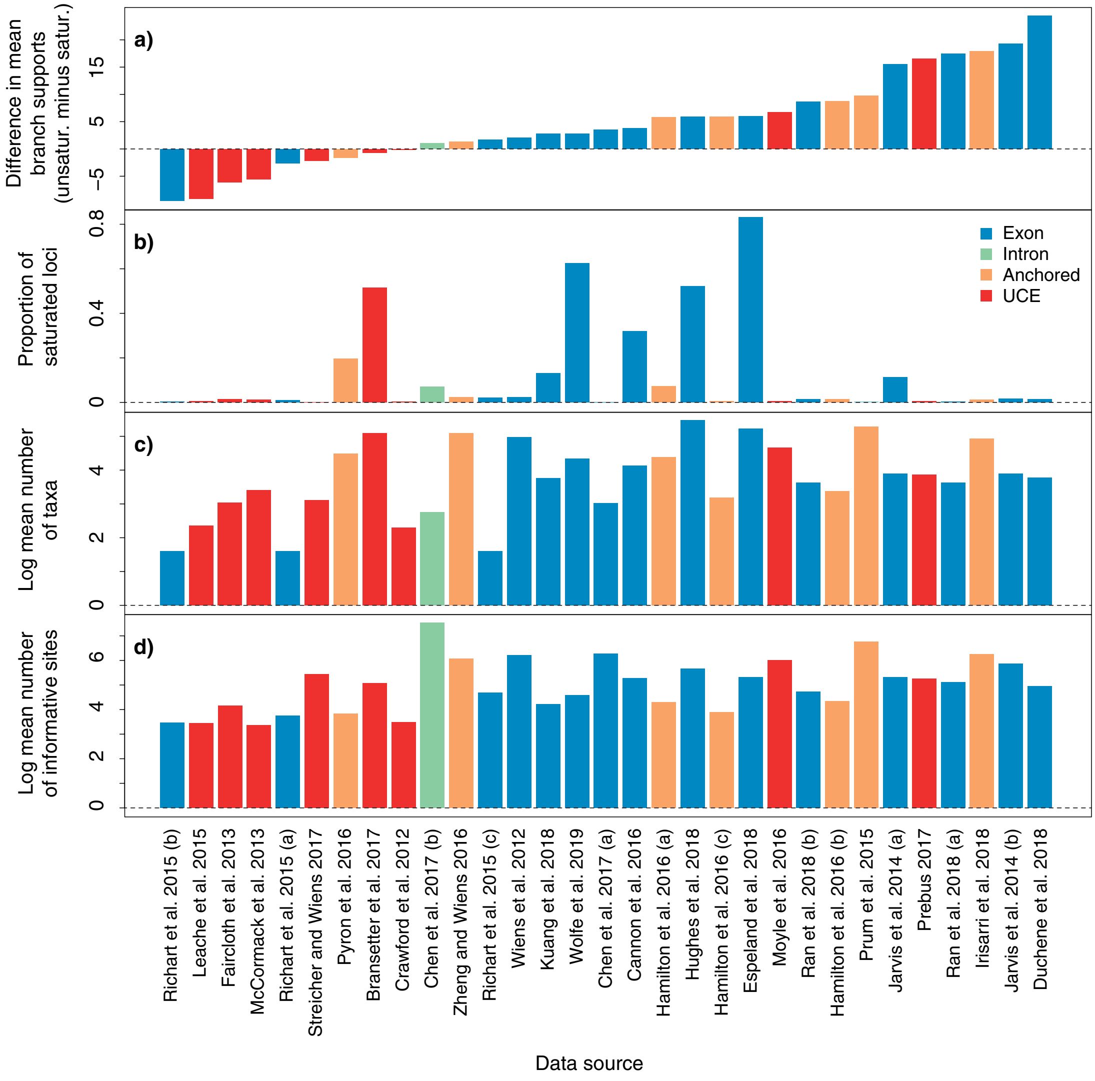

FIGURE 6. Characteristics of empirical phylogenomic data sets ranked by the difference in mean branch support between gene trees from unsaturated versus saturated loci (a). Mean branch supports are taken from mean aLRT supports (Guindon et al. 2010) across branches. The portion of saturated loci (b) is 1 minus the portion of unsaturated loci. Values of number of taxa and number of informative sites are the mean across loci in each data set.

data sets. Assessing saturation in alignments as a whole can be particularly useful after filtering by taxa (e.g., Aberer et al. 2013; Mai and Mirarab 2018) and by sites (e.g., Ranwez et al. 2011; Whelan et al. 2018). Using a combination of methods for filtering by rate might also reduce the portion of the false negatives that arise in tests of historical signal due to fast-evolving segments of alignments (Dornburg et al. 2019).

Our findings on the performance of the entropy saturation test also point to the importance of assessing substitution model adequacy in phylogenetic studies (e.g., Goldman 1993; Bollback 2002; Weiss and von Haeseler 2003; Foster 2004; Brown 2014; Duchêne et al. 2018c). Our results from the saturation tests applied to all sites of an alignment suggest that complex evolutionary models can mislead any entropy-based saturation test on an alignment as a whole. Examples of more complex models that can mislead the test include a process in which substitution is dominated by a small number of distinct categories of base frequencies, also known for its implementation as the CAT model (Fitch and Markowitz 1970; Lartillot and Philippe 2004). Another difficult scenario is that of the covarion substitution process, where substitution types are constrained at various points in evolutionary time (Miyamoto and Fitch 1996). These models do not lead to alignment-wide base frequencies under maximum entropy, such that our saturation test is an inadequate representation of the model that generated the data. Therefore, we encourage the testing of a broad range of evolutionary models in phylogenomics. Alternatively, researchers can use multiple methods of model assessment, beginning with tests of signals in individual taxa or sites, followed by tests of model adequacy, and finishing with overall

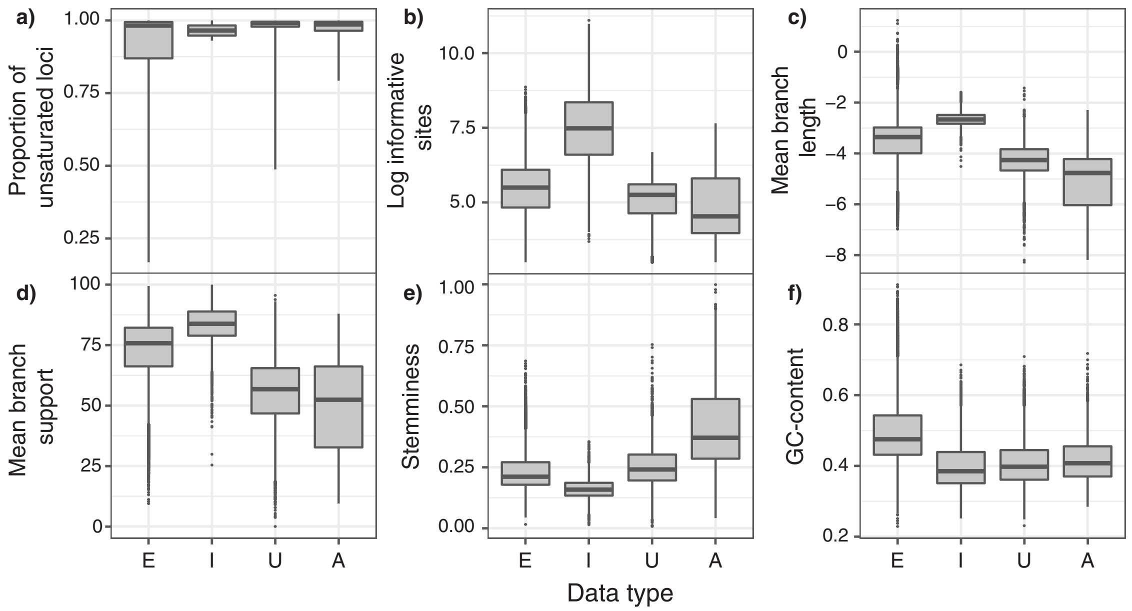

FIGURE 7. Characteristics of each data type grouped across the studies examined. Data types include exon (E), intron (I), ultraconserved elements (U), and anchored-enriched regions (A). Characteristics of data types include (a) the results from the new test of saturation as the portion of loci across studies identified as unsaturated. Data in panel (a) represent values per data set, while those in panels (b–f) show data per locus. The log number of informative sites (b) refers to sites that are parsimony-informative. Mean branch supports (c) were calculated using aLRT (Guindon et al. 2010). Stemminess (e) refers to the ratio of summed internal to external branch lengths (Fiala and Sokal 1985).

tests of historical signal (e.g., Dávalos and Perkins 2008; Liu et al. 2014).

The entropy test of substitution saturation on phylogenetically informative sites shows good performance across a broad range of scenarios and is likely to be robust to other factors that were not explored in this study. For example, the test is likely to perform well in the presence of rate variation across sites, provided that this form of variation is modeled appropriately (Kalyaanamoorthy et al. 2017). However, due to the unusual nature of the tree-topology parameter, tests of data quality are generally dissociated from the actual performance of phylogenetic methods (also see Duchêne et al. 2017, 2018c). Outlier fast-evolving lineages might pose a challenge to tests of historical signal like the entropy saturation test, due to the possible mixed signals of closely related and highly divergent taxa (Dornburg et al. 2019). Therefore, it is useful to complement tests of the historical signal with tests of the plausibility of branch-length estimates, or of the consistency in phylogenetic signal across an alignment (e.g., Minin et al. 2003; Aberer et al. 2013; Mai and Mirarab 2018).

An important matter when developing tests of phylogenetic signal or model adequacy is to identify appropriate critical values that balance the rates of true and false positives. Identifying such critical values can be a difficult task, for many reasons. In particular, phylogenetic information is not straightforward to capture in test statistics that summarize the features of sequence alignments (Duchêne et al. 2018c). We find that the entropy saturation test is associated with several factors that affect the quality of phylogenetic inference, such as stemminess and evolutionary rate. However, the test might vary in usefulness across data sets and among loci within a phylogenomic data set. This is in part because of variance across loci in the performance of the substitution model, which will affect the performance of the test of saturation. Users of tests of model adequacy and saturation need to be aware of this limitation of the tests, and we recommend at least reporting the predicted rates of true and false positives across various data sizes from our simulations. Further work on methods of reporting the uncertainty around critical values of assessment will be valuable.

Alternative methods of assessing substitution saturation might prove to have better performance. For example, a common practice for model assessment in evolutionary biology is to use null distributions based on simulations (Brown and Thomson 2018), which allows for a test that is highly tailored to the data. However, simulations can be computationally demanding, and a simulations-based test would be dependent on using an adequate substitution model. Yet another alternative approach is to train a machine-learning algorithm to assess historical signal. Machine learning has recently been proposed in phylogenetics for substitution model selection (Abadi et al. 2020), inference of tree topology (Suvorov et al. 2019), species delimitation (Derkarabetian et al. 2019), and analyses of molecular rates across lineages (Tao et al. 2019). A random forest or an artificial neural network might prove to be highly effective for identifying the factors that are associated with accurate inferences. Similarly, these algorithms could be trained for identifying sequence alignments with misleading signals. These alternative methods might have superior performance to entropy-based tests of saturation. Nonetheless, the computational demand of the test presented here is minimal and is unlikely to be reduced substantially using other frameworks.

Substitution saturation is detrimental to phylogenetic inference and is common in phylogenomic data sets, but it can be effectively identified using appropriate tests. Phylogenomic data sets are now widespread and researchers need to identify the data, models, and methods that are most suitable for answering the biological questions being posed (e.g., Reddy et al. 2017; Molloy and Warnow 2018; Richards et al. 2018; Bravo et al. 2019; Karin et al. 2019; Duchêne et al. 2020). The entropy test performs well across a wide range of simulation scenarios, and we provide guidelines for its usage. Tests of substitution saturation and model adequacy will improve the quality of phylogenetic inference in the genomic era, particularly in studies using data from exons or introns.

Supplementary Material

Data available from the Zenodo digital repository: https://doi.org/10.5281/zenodo.5131558.

Acknowledgments

The authors acknowledge the Sydney Informatics Hub and the University of Sydney's high-performance computing cluster Artemis for providing computing resources that contributed to the research results reported in this paper.

Funding

This work was supported by funding from the Carlsbergfondet of Denmark (CF18-0223 to D.A.D.) and (DP160104173 and FT160100167 to S.Y.W.H.).

References

Abadi S., Avram O., Rosset S., Pupko T., Mayrose I. 2020. ModelTeller: Model selection for optimal phylogenetic reconstruction using machine learning. Mol. Biol. Evol. 37:3338–3352.

Aberer A.J., Krompass D., Stamatakis A. 2013. Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. Syst. Biol. 62:162–166.

Bollback J.P. 2002. Bayesian model adequacy and choice in phylogenetics. Mol. Biol. Evol. 19:1171–1180.

Branstetter M.G., Danforth B.N., Pitts J.P., Faircloth B.C., Ward P.S., Buffington M.L., Gates M.W., Kula R.R., Brady S.G. 2017. Phylogenomic insights into the evolution of stinging wasps and the origins of ants and bees. Curr. Biol. 27:1019–1025.

Bravo G.A., Antonelli A., Bacon C.D., Bartoszek K., Blom M.P.K., Huynh S., Jones G., Lacey Knowles L., Lamichhaney S., Marcussen T., Morlon H., Nakhleh L.K., Oxelman B., Pfeil B., Schliep A., Wahlberg N., Werneck F.P., Wiedenhoeft J., Willows-Munro S., Edwards S. V. 2019. Embracing heterogeneity: coalescing the tree of life and the future of phylogenomics. PeerJ. 6:e26449v3.

Brown J.M. 2014. Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. Syst. Biol. 63:334–348.

Brown J.M., Thomson R.C. 2018. Evaluating model performance in evolutionary biology. Annu. Rev. Ecol. Evol. Syst. 49:95–114.

Brown W.M., Prager E.M., Wang A., Wilson A.C. 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. J. Mol. Evol. 18:225–239.

Cannon J.T., Vellutini B.C., Smith J., Ronquist F., Jondelius U., Hejnol A. 2016. Xenacoelomorpha is the sister group to Nephrozoa. Nature. 530:89–93.

Chen M.-Y., Liang D., Zhang P. 2017. Phylogenomic resolution of the phylogeny of laurasiatherian mammals: exploring phylogenetic signals within coding and noncoding sequences. Genome Biol. Evol. 9:1998–2012.

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.

Dávalos L.M., Perkins S.L. 2008. Saturation and base composition bias explain phylogenomic conflict in *Plasmodium*. Genomics. 91:433–442.

Derkarabetian S., Castillo S., Koo P.K., Ovchinnikov S., Hedin M. 2019. A demonstration of unsupervised machine learning in species delimitation. Mol. Phylogenet. Evol. 139:106562.

Dornburg A., Su Z., Townsend J.P. 2019. Optimal rates for phylogenetic inference and experimental design in the era of genome-scale data sets. Syst. Biol. 68:145–156.

Duchêne D.A., Bragg J.G., Duchêne S., Neaves L.E., Potter S., Moritz C., Johnson R.N., Ho S.Y.W., Eldridge M.D.B. 2018a. Analysis of phylogenomic tree space resolves relationships among marsupial families. Syst. Biol. 67:400–412.

Duchêne D.A., Duchêne S., Ho S.Y.W. 2017. New statistical criteria detect phylogenetic bias caused by compositional heterogeneity. Mol. Biol. Evol. 34:1529–1534.

Duchêne D.A., Duchêne S., Ho S.Y.W. 2018b. PhyloMAd: efficient assessment of phylogenomic model adequacy. Bioinformatics. 34:2300–2301.

Duchêne D.A., Duchêne S., Ho S.Y.W. 2018c. Differences in performance among test statistics for assessing phylogenomic model adequacy. Genome Biol. Evol. 10:1375–1388.

Duchêne D.A., Tong K.J., Foster C.S., Duchêne S., Lanfear R., Ho S.Y.W. 2020. Linking branch lengths across sets of loci provides the highest statistical support for phylogenetic inference. Mol. Biol. Evol. 37:1202–1210.

Espeland M., Breinholt J., Willmott K.R., Warren A.D., Vila R., Toussaint E.F.A., Maunsell S.C., Aduse-Poku K., Talavera G., Eastwood R. 2018. A comprehensive and dated phylogenomic analysis of butterflies. Curr. Biol. 28:770–778.

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. 61:717–26.

Faircloth B.C., Sorenson L., Santini F., Alfaro M.E. 2013. A phylogenomic perspective on the radiation of ray-finned fishes based upon targeted sequencing of ultraconserved elements (UCEs). PLoS One. 8:e65923.

Fiala K.L., Sokal R.R. 1985. Factors determining the accuracy of cladogram estimation: evaluation using computer simulation. Evolution. 39:609–622.

Fitch W.M., Markowitz E. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genet. 4:579–593.

Foster P.G. 2004. Modeling compositional heterogeneity. Syst. Biol. 53:485–495.

Goldman N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. 36:182–198.

Goldman N. 1998. Phylogenetic information and experimental design in molecular systematics. Proc. R. Soc. B Biol. Sci. 265: 1779–1786.

Guindon S., Dufayard J.-F., Lefort V., Anisimova M., Hordijk W., Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0.

Syst. Biol. 59:307–321.

Hamilton C.A., Lemmon A.R., Lemmon E.M., Bond J.E. 2016. Expanding anchored hybrid enrichment to resolve both deep and shallow relationships within the spider tree of life. BMC Evol. Biol. 16:212.

Han H.-Y., Ro K.-E. 2005. Molecular phylogeny of the superfamily Tephritoidea (Insecta: Diptera): new evidence from the mitochondrial 12S, 16S, and COII genes. Mol. Phylogenet. Evol. 34:416–430.

Ho S.Y.W., Jermiin L. 2004. Tracing the decay of the historical signal in biological sequence data. Syst. Biol. 53:623–637.

Hughes L.C., Ortí G., Huang Y., Sun Y., Baldwin C.C., Thompson A.W., Arcila D., Betancur R., Li C., Becker L., Bellora N., Zhao X., Li X., Wang M., Fang C., Xie B., Zhoui Z., Huang H., Chen S., Venkatesh B., Shi Q. 2018. Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. Proc. Natl. Acad. Sci. USA. 115:6249–6254.

Irisarri I., Singh P., Koblmüller S., Torres-Dowdall J., Henning F., Franchini P., Fischer C., Lemmon A.R., Lemmon E.M., Thallinger G.G., Sturmbauer C., Meyer A. 2018. Phylogenomics uncovers early hybridization and adaptive loci shaping the radiation of Lake Tanganyika cichlid fishes. Nat. Commun. 9:3159.

Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y.W., Faircloth B.C., Nabholz B., Howard J.T., Suh A., Weber C.C., da Fonseca R.R., Li J., Zhang F., Li H., Zhou L., Narula N., Liu L., Ganapathy G., Boussau B., Bayzid M.S., Zavidovych V., Subramanian S., Gabaldon T., Capella-Gutierrez S., Huerta-Cepas J., Rekepalli B., Munch K., Schierup M., Lindow B., Warren W.C., Ray D., Green R.E., Bruford M.W., Zhan X., Dixon A., Li S., Li N., Huang Y., Derryberry E.P., Bertelsen M.F., Sheldon F.H., Brumfield R.T., Mello C. V., Lovell P. V., Wirthlin M., Schneider M.P.C., Prosdocimi F., Samaniego J.A., Velazquez A.M. V., Alfaro-Nunez A., Campos P.F., Petersen B., Sicheritz-Ponten T., Pas A., Bailey T., Scofield P., Bunce M., Lambert D.M., Zhou Q., Perelman P., Driskell A.C., Shapiro B., Xiong Z., Zeng Y., Liu S., Li Z., Liu B., Wu K., Xiao J., Yinqi X., Zheng Q., Zhang Y., Yang H., Wang J., Smeds L., Rheindt F.E., Braun M., Fjeldsa J., Orlando L., Barker F.K., Jonsson K.A., Johnson W., Koepfli K.-P., O'Brien S., Haussler D., Ryder O.A., Rahbek C., Willerslev E., Graves G.R., Glenn T.C., McCormack J., Burt D., Ellegren H., Alstrom P., Edwards S. V., Stamatakis A., Mindell D.P., Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T.P., Zhang G. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science. 346:1320–1331.

Kalyaanamoorthy S., Minh B.Q., Wong T.K.F., von Haeseler A., Jermiin L.S. 2017. ModelFinder: Fast model selection for accurate phylogenetic estimates. Nat. Methods. 14:587–589.

Karin B.R., Gamble T., Jackman T.R. 2019. Optimizing phylogenomics with rapidly evolving long exons: comparison with anchored hybrid enrichment and ultraconserved elements. Mol. Biol. Evol. 37:904–922.

Klopfstein S., Massingham T., Goldman N. 2017. More on the best evolutionary rate for phylogenetic analysis. Syst. Biol. 66:769–785.

Kuang T., Tornabene L., Li J., Jiang J., Chakrabarty P., Sparks J.S., Naylor G.J.P., Li C. 2018. Phylogenomic analysis on the exceptionally diverse fish clade Gobioidei (Actinopterygii: Gobiiformes) and data-filtering based on molecular clocklikeness. Mol. Phylogenet. Evol. 128:192–202.

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.

Leaché A.D., Chavez A.S., Jones L.N., Grummer J.A., Gottscho A.D., Linkem C.W. 2015. Phylogenomics of phrynosomatid lizards: conflicting signals from sequence capture versus restriction site associated DNA sequencing. Genome Biol. Evol. 7:706–719.

Lemmon A.R., Emme S.A., Lemmon E.M. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. 61:727–744.

Léveillé-Bourret É., Starr J.R., Ford B.A., Moriarty Lemmon E., Lemmon A.R. 2018. Resolving rapid radiations within angiosperm families using anchored phylogenomics. Syst. Biol. 67:94–112.

Liu Y., Cox C.J., Wang W., Goffinet B. 2014. Mitochondrial phylogenomics of early land plants: mitigating the effects of saturation, compositional heterogeneity, and codon-usage bias. Syst. Biol. 63:862–878.

Liu Y., Johnson M.G., Cox C.J., Medina R., Devos N., Vanderpoorten A., Hedenäs L., Bell N.E., Shevock J.R., Aguero B., Quandt D., Wickett N.J., Shaw A.J., Goffinet B. 2019. Resolution of the ordinal phylogeny of mosses using targeted exons from organellar and nuclear genomes. Nat. Commun. 10:1485.

Mai U., Mirarab S. 2018. TreeShrink: Fast and accurate detection of outlier long branches in collections of phylogenetic trees. BMC Genomics. 19:272.

McCormack J.E., Harvey M.G., Faircloth B.C., Crawford N.G., Glenn T.C., Brumfield R.T. 2013. A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. PLoS One. 8:e54848.

Mindell D.P., Honeycutt R.L. 1990. Ribosomal RNA in vertebrates: evolution and phylogenetic applications. Annu. Rev. Ecol. Syst. 21:541–566.

Minin V., Abdo Z., Joyce P., Sullivan J. 2003. Performance-based selection of likelihood models for phylogeny estimation. Syst. Biol. 52:674–683.

Miyamoto M.M., Fitch W.M. 1996. Constraints on protein evolution and the age of the eubacteria/eukaryote split. Syst. Biol. 45:568–575.

Molloy E.K., Warnow T. 2018. To include or not to include: the impact of gene filtering on species tree estimation methods. Syst. Biol. 67:285–303.

Moyle R.G., Oliveros C.H., Andersen M.J., Hosner P.A., Benz B.W., Manthey J.D., Travers S.L., Brown R.M., Faircloth B.C. 2016. Tectonic collision and uplift of Wallacea triggered the global songbird radiation. Nat. Commun. 7:12709.

Nguyen L.-T., Schmidt H.A., von Haeseler A., Minh B.Q. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32:268–274.

Penny D., Hendy M.D. 1985. The use of tree comparison metrics. Syst. Zool. 34:75–82.

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., Forterre P. 1999. The rooting of the universal tree of life is not reliable. J. Mol. Evol. 49:509–523.

Philippe H., Lecointre G., Le H., Le Guyader H. 1996. A critical study of homoplasy in molecular data with the use of a morpholoogically based cladogram, and its consequences for character weighting. Mol. Biol. Evol. 13:1174–1186.

Prebus M. 2017. Insights into the evolution, biogeography and natural history of the acorn ants, genus *Temnothorax* Mayr (Hymenoptera: Formicidae). BMC Evol. Biol. 17:250.

Prum R.O., Berv J.S., Dornburg A., Field D.J., Townsend J.P., Lemmon E.M., Lemmon A.R. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature. 526:569–573.

Pyron R.A., Hsieh F.W., Lemmon A.R., Lemmon E.M., Hendry C.R. 2016. Integrating phylogenomic and morphological data to assess candidate species-delimitation models in brown and red-bellied snakes (*Storeria*). Zool. J. Linn. Soc. 177:937–949.

Ran J.H., Shen T.T., Wang M.M., Wang X.Q. 2018. Phylogenomics resolves the deep phylogeny of seed plants and indicates partial convergent or homoplastic evolution between Gnetales and angiosperms. Proc. R. Soc. B Biol. Sci. 285:20181012.

Ranwez V., Harispe S., Delsuc F., Douzery E.J.P. 2011. MACSE: Multiple alignment of coding sequences accounting for frameshifts and stop codons. PLoS One. 6:e22594.

Reddy S., Kimball R.T., Pandey A., Hosner P.A., Braun M.J., Hackett S.J., Han K.-L., Harshman J., Huddleston C.J., Kingston S., Marks B.D., Miglia K.J., Moore W.S., Sheldon F.H., Witt C.C., Yuri T., Braun E.L. 2017. Why do phylogenomic data sets yield conflicting trees? Data type influences the avian Tree of Life more than taxon sampling. Syst. Biol. 66:857–879.

Revell L., Harmon L., Glor R. 2005. Under-parameterized model of sequence evolution leads to bias in the estimation of diversification rates from molecular phylogenies. Syst. Biol. 54:973–983.

Richards E.J., Brown J.M., Barley A.J., Chong R.A., Thomson R.C. 2018. Variation across mitochondrial gene trees provides evidence for systematic error: how much gene tree variation Is biological? Syst. Biol. 67:847–860.

Richart C.H., Hayashi C.Y., Hedin M. 2016. Phylogenomic analyses resolve an ancient trichotomy at the base of Ischyropsalidoidea (Arachnida, Opiliones) despite high levels of gene tree conflict and unequal minority resolution frequencies. Mol. Phylogenet. Evol. 95:171–182.

Robinson D.F., Foulds L.R. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.

Shen H., Jin D., Shu J.-P., Zhou X.-L., Lei M., Wei R., Shang H., Wei H.-J., Zhang R., Liu L., Gu Y.-F., Zhang X.-C., Yan Y.-H. 2018. Large-scale phylogenomic analysis resolves a backbone phylogeny in ferns. Gigascience. 7:gix116.

Streicher J.W., Wiens J.J. 2017. Phylogenomic analyses of more than 4000 nuclear loci resolve the origin of snakes among lizard families. Biol. Lett. 13:20170393.

Strimmer K., Von Haeseler A. 1997. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. Proc. Natl. Acad. Sci. USA. 94:6815–6819.

Sullivan J., Joyce P. 2005. Model selection in phylogenetics. Annu. Rev. Ecol. Evol. Syst. 36:445–466.

Susko E., Roger A.J. 2012. The probability of correctly resolving a split as an experimental design criterion in phylogenetics. Syst. Biol. 61:811–821.

Suvorov A., Hochuli J., Schrider D.R. 2019. Accurate inference of tree topologies from multiple sequence alignments using deep learning. Syst. Biol. 69:221–233.

Tao Q., Tamura K., U. Battistuzzi F., Kumar S. 2019. A machine learning method for detecting autocorrelation of evolutionary rates in large phylogenies. Mol. Biol. Evol. 36:811–824.

Townsend J.P. 2007. Profiling phylogenetic informativeness. Syst. Biol. 56:222–231.

Townsend J.P., Su Z., Tekle Y.I. 2012. Phylogenetic signal and noise: predicting the power of a data set to resolve phylogeny. Syst. Biol. 61:835–849.

Weiss G., von Haeseler A. 2003. Testing substitution models within a phylogenetic tree. Mol. Biol. Evol. 20:572–578.

Whelan S., Irisarri I., Burki F. 2018. PREQUAL: detecting non-homologous characters in sets of unaligned homologous sequences. Bioinformatics. 34:3929–3930.

Wiens J.J., Hutter C.R., Mulcahy D.G., Noonan B.P., Townsend T.M., Sites J.W., Reeder T.W. 2012. Resolving the phylogeny of lizards and snakes (Squamata) with extensive sampling of genes and species. Biol. Lett. 8:1043–1046.

Wolfe J.M., Breinholt J.W., Crandall K.A., Lemmon A.R., Lemmon E.M., Timm L.E., Siddall M.E., Bracken-Grissom H.D. 2019. A phylogenomic framework, evolutionary timeline and genomic resources for comparative studies of decapod crustaceans. Proc. R. Soc. B Biol. Sci. 286:20190079.

Xia X., Xie Z., Salemi M., Chen L., Wang Y. 2003. An index of substitution saturation and its application. Mol. Phylogenet. Evol. 26:1–7.

Yang Z. 1998. On the best evolutionary rate for phylogenetic analysis. Syst. Biol. 47:125–133.

Zheng Y., Wiens J.J. 2016. Combining phylogenomic and supermatrix approaches, and a time-calibrated phylogeny for squamate reptiles (lizards and snakes) based on 52 genes and 4162 species. Mol. Phylogenet. Evol. 94:537–547.