Syst. Biol. 69(2):308–324, 2020

DOI:10.1093/sysbio/syz058
Advance Access publication August 29, 2019

Quartet-Based Computations of Internode Certainty Provide Robust Measures of Phylogenetic Incongruence

XIAOFAN ZHOU[1], SARAH LUTTEROPP[2], LUCAS CZECH[2], ALEXANDROS STAMATAKIS[2,3], MORITZ VON LOOZ[3], AND ANTONIS ROKAS[4,*]

[1]*Guangdong Province Key Laboratory of Microbial Signals and Disease Control, Integrative Microbiology Research Centre, South China Agricultural University, 483 Wushan Road, Guangzhou 510642, P.R. China;* [2]*The Exelixis Lab, Scientific Computing Group, Heidelberg Institute for Theoretical Studies, Schloss-Wolfsbrunnenweg 35, Heidelberg D-68159, Germany;* [3]*Department of Informatics, Institute of Theoretical Informatics, Karlsruhe Institute of Technology, Postfach 6980, 76128 Karlsruhe, Germany; and* [4]*Department of Biological Sciences, Vanderbilt University, VU Station B #35-1634, Nashville, TN 37235, USA*

**Correspondence to be sent to: Department of Biological Sciences, Vanderbilt University, VU Station B #35-1634, Nashville, TN 37235, USA; E-mail: antonis.rokas@vanderbilt.edu.*

Received 3 August 2017; reviews returned 29 July 2019; accepted 26 August 2019
Associate Editor: Olivie Gascuel

Abstract.—Incongruence, or topological conflict, is prevalent in genome-scale data sets. Internode certainty (IC) and related measures were recently introduced to explicitly quantify the level of incongruence of a given internal branch among a set of phylogenetic trees and complement regular branch support measures (e.g., bootstrap, posterior probability) that instead assess the statistical confidence of inference. Since most phylogenomic studies contain data partitions (e.g., genes) with missing taxa and IC scores stem from the frequencies of bipartitions (or splits) on a set of trees, IC score calculation typically requires adjusting the frequencies of bipartitions from these partial gene trees. However, when the proportion of missing taxa is high, the scores yielded by current approaches that adjust bipartition frequencies in partial gene trees differ substantially from each other and tend to be overestimates. To overcome these issues, we developed three new IC measures based on the frequencies of quartets, which naturally apply to both complete and partial trees. Comparison of our new quartet-based measures to previous bipartition-based measures on simulated data shows that: (1) on complete data sets, both quartet-based and bipartition-based measures yield very similar IC scores; (2) IC scores of quartet-based measures on a given data set with and without missing taxa are more similar than the scores of bipartition-based measures; and (3) quartet-based measures are more robust to the absence of phylogenetic signal and errors in phylogenetic inference than bipartition-based measures. Additionally, the analysis of an empirical mammalian phylogenomic data set using our quartet-based measures reveals the presence of substantial levels of incongruence for numerous internal branches. An efficient open-source implementation of these quartet-based measures is freely available in the program *QuartetScores* (https://github.com/lutteropp/QuartetScores). [Missing taxa; phylogenetics; phylogenomics; phylogenetic conflict; phylogenetic discordance; phylogenetic signal; robustness.]

Recent advances in DNA sequencing technologies have greatly facilitated the generation of genome-scale data for phylogenetic inference in diverse groups of organisms, including fungi (e.g., Nagy et al. 2014; Shen et al. 2016; Shen et al. 2018; Steenwyk et al. 2019), plants (e.g., Wickett et al. 2014; Yang et al. 2015), and animals (e.g., Song et al. 2012; Jarvis et al. 2014). Incongruence (i.e., the presence of topological conflict) between individual gene trees in each one of these phylogenomic data matrices is the rule rather than the exception. The hundreds or thousands of genes examined in a study each yield their own distinct topologies (e.g., Song et al. 2012; Salichos and Rokas 2013; Zhong et al. 2013). The observed incongruence can be partly attributed to gene tree estimation errors caused by analytical reasons including insufficient information in the data, misspecification of evolutionary models, or inadequate tree search (Jeffroy et al. 2006; Kumar et al. 2012). On the other hand, the evolutionary histories of genes can also be genuinely different from each other and from the underlying species phylogeny due to biological processes such as incomplete lineage sorting (ILS), introgression, hybridization, natural selection, and horizontal gene transfer (Maddison 1997; Slowinski and Page 1999; Castoe et al. 2009; Degnan and Rosenberg 2009).

Given the prevalence of phylogenetic incongruence, its unequal distribution across branches of a phylogeny, and its key role in assessing the robustness of species tree inference (Salichos and Rokas 2013), it is important that our measures of incongruence are accurate. Salichos et al. recently developed several novel information theory-based measures to quantify incongruence among a set of "evaluation" trees (e.g., gene trees) with respect to the internal branches (which they referred to as "internodes") in a "reference" tree (e.g., the species tree) (Salichos and Rokas 2013; Salichos et al. 2014). In brief, for the bipartition defined by a given internal branch in the reference tree, its conflicting bipartitions are initially extracted from the evaluation tree set. Then, Shannon's entropy (Shannon 1948) is calculated from the frequencies of occurrence (in the evaluation trees) of both the reference bipartition and the conflicting ones. In this way, the diversity and strengths of conflicting signals are integrated into a single measure of the degree of certainty (or uncertainty) about the phylogenetic relationship defined by the internal branch in the reference tree. Measurement of internode certainty (IC) comes in two flavors; the IC score only takes into account the reference bipartition and the most prevalent conflicting bipartition, while the IC all (ICA) score also considers all other conflicting bipartitions that are sufficiently

TABLE 1. Definition of acronyms used in this paper

Acronym	Definition
IC(A)	Internode certainty (all)
LIC(A)	Lossless internode certainty (all)
PIC(A)	Probabilistic internode certainty (all)
Q-IC	Quartet internode certainty
LQ-IC	Lowest quartet internode certainty
(E)QP-IC	(Extended) Quadripartition internode certainty
BS	Bootstrap
GSF	Gens support frequency
LPP	Local posterior probability
QS	Quartet sampling
ILS	Incomplete lineage sorting
MSC	Multi-species coalescent

frequent (see Table 1 for the list of all acronyms used in this study).

The original IC/ICA scores are applicable only if all evaluation trees are complete, that is, they contain exactly the same taxa as the reference tree (Salichos et al. 2014). However, in phylogenomic studies, it is common that the sequences of many (or even most) genes are only available from taxon subsets, giving rise to partial gene trees. To meet the need to quantify incongruence in evaluation tree sets that contain partial trees, Kobert et al. (2016) developed mathematical approaches to adjust the frequencies of bipartitions from partial trees in the calculation of IC/ICA scores. Specifically, the authors developed three adjustment schemes that differ on how the frequency of a bipartition with missing taxa is corrected (Kobert et al. 2016): (1) *Probabilistic*—the frequency of the incomplete bipartition is distributed equally to all possible complete bipartitions (i.e., containing all taxa) that are compatible with it; (2) *Observed*—the frequency of the incomplete bipartition is distributed equally to only those compatible, complete bipartitions observed in the reference and evaluation trees; and (3) *Lossless*—similar to *Observed*, but with the restriction that the complete bipartitions also have to be mutually conflicting. An approach similar to the *Lossless* adjustment scheme was also independently developed by Smith et al. (2015).

IC and related measures are valuable and effective tools in identifying phylogenetic incongruence and have been quickly adopted in phylogenomic studies (Chen et al. 2015; Wang et al. 2015; Li et al. 2016; Shen et al. 2016; Chesters 2017; Krabberod et al. 2017; Leveille-Bourret et al. 2017; Fernandez et al. 2018; Shen et al. 2018; Steenwyk et al. 2019; Yang et al. 2018), yet they still exhibit several practical and theoretical limitations. On the practical side, for data sets with high proportions of missing taxa (e.g., all genes trees are partial), the aforementioned adjustment schemes can considerably overestimate IC/ICA scores (Kobert et al. 2016). Additionally, alternative adjustment schemes might generate substantially different scores (Kobert et al. 2016) and it is often unclear which scheme is better. On the theoretical side, for the ICA measure, the exact number of conflicting bipartitions to be considered can only be determined *post hoc* from the evaluation trees (Salichos et al. 2014; Kobert et al. 2016), which might lead to unexpected behavior. To illustrate this point, consider the following example with one reference bipartition and two conflicting bipartitions. If we set their frequencies to 80%:10%:10%, 80%:15%:5%, and 80%:19%:1%, the ICA scores would be 0.42, 0.44, and 0.51, respectively. That is, the ICA score of the internal branch increases as one of the conflicting bipartitions appears more frequently. However, if all 20% of the conflicting signal stems entirely from one bipartition (i.e., 80%:20%), then the ICA score drops again to 0.28. This is because the ICA score calculation now involves only two bipartitions instead of three, which changes the base of logarithm in Shannon's entropy equation (Shannon 1948) from 3 to 2, thereby drastically lowering the score.

One potential solution to these practical and theoretical issues is to base the quantification of phylogenetic incongruence on quartets instead of bipartitions (see also Pease et al. 2018). Quartets (i.e., sets of four taxa) are the most basic unit of information in unrooted phylogenetic trees and have long been used in molecular phylogenetics for a wide range of purposes, including tree reconstruction (Strimmer and von Haeseler 1996; Chifman and Kubatko 2014; Avni et al. 2015; Mirarab and Warnow 2015) and rogue taxon identification (Wilkinson 2006; Aberer and Stamatakis 2011). Several properties make quartets particularly attractive for quantifying IC. First, both the reference and evaluation trees can be decomposed into sets of induced quartets. Second, the quartet set of the reference tree is a superset of the quartet set of every evaluation tree. Therefore, both complete and partial evaluation trees can be naturally compared with the reference tree at the quartet level without any further need for adjustment. In addition, evaluation trees with more missing taxa will contribute fewer quartets to the quantification (since the number of quartets contained in a tree grows polynomially with the number of taxa), providing a natural way to weigh evaluation trees of different sizes. Moreover, every quartet tree has a fixed number of three alternative topologies, hence two conflicting topologies will always be expected for every quartet topology in the reference tree regardless of the taxa present in the evaluation trees.

Quartets have been previously used to detect conflicting signal in phylogenetic data sets. For instance, Driskell et al. (2004) used quartets to identify informative gene trees for subsequent comparison with reference species trees. Moreover, the quartet-mapping approach assesses the phylogenetic content of a multiple sequence alignment by analyzing every possible combination of four sequences (Strimmer and von Haeseler 1997; Nieselt-Struwe and von Haeseler 2001). The approach was later adopted to study the incongruence among quartet gene trees for sets of four species (Nesbo et al. 2001; Zhaxybayeva and Gogarten 2002; Zhaxybayeva et al. 2006). However, these approaches are limited to analyzing sets of four taxa, instead of internal branches in a species tree.

Here, we introduce three new quartet-based measures for quantifying incongruence among phylogenetic trees, which can be calculated using the freely available program https://github.com/lutteropp/QuartetScores. Much like existing bipartition-based IC measures (Salichos et al. 2014; Kobert et al. 2016), the output of all three new measures are IC scores for all internal branches in the reference tree, which reflect the degree of certainty of the bipartition defined by each internal branch. Using both simulated and biological data sets, we show that quartet-based and bipartition-based measures perform equally well in calculating IC on sets of complete trees and that quartet-based measures outperform bipartition-based ones on sets that contain missing data. Additionally, we establish the sensitivity of quartet-based IC measures to specific analytic challenges, such as the lack of phylogenetic signal and topological errors in reference trees. Application of these new measures on an empirical mammalian phylogenomic data set reveals high degrees of phylogenetic incongruence for certain internal branches that are consistent with the current understanding of contentious relationships in mammals. Overall, our results suggest that our newly developed quartet-based measures are useful for more accurately quantifying phylogenetic incongruence.

Three New Quartet-Based Measures for Estimating Internode Certainty

All three quartet-based measures require as input a reference tree T_R and a set of evaluation trees T_E; only unrooted trees are considered. The taxon set of the reference tree $S(T_R)$ should be equal to the union of the taxon sets of all evaluation trees $S(T_E)$. All evaluation trees may have the same taxon set as $S(T_R)$ (e.g., T_R and T_E are the bootstrap consensus tree and bootstrap replicate trees, respectively, from a single-gene phylogenetic analysis). Alternatively, the taxon sets of some or all evaluation trees may be strict subsets of $S(T_R)$ (e.g., T_R and T_E are the coalescent-based species tree and single-gene trees, respectively, from a phylogenomic analysis where some genes are missing from some taxa).

All three measures require the generation of a list of quartets induced by T_R and the occurrences of their alternative topologies in T_E (Fig. 1a). Unresolved quartet topologies in polytomous evaluation trees are discarded. The three measures differ in whether all (or some) possible quartets are used and how they are used, which in turn influences how the IC is calculated for each internal branch in T_R (Fig. 1b–e). To illustrate the calculation of our three quartet-based IC scores, we use an example data set consisting of a six-species reference tree T_R and an evaluation tree set T_E that includes one complete tree topology and three partial tree topologies, each of which appears a given number of times in T_E (shown along the respective tree topology; Fig. 1a). In this example, we focus on calculating the quartet-based IC scores for the internal branch separating (A, B) from (C, D, E, F).

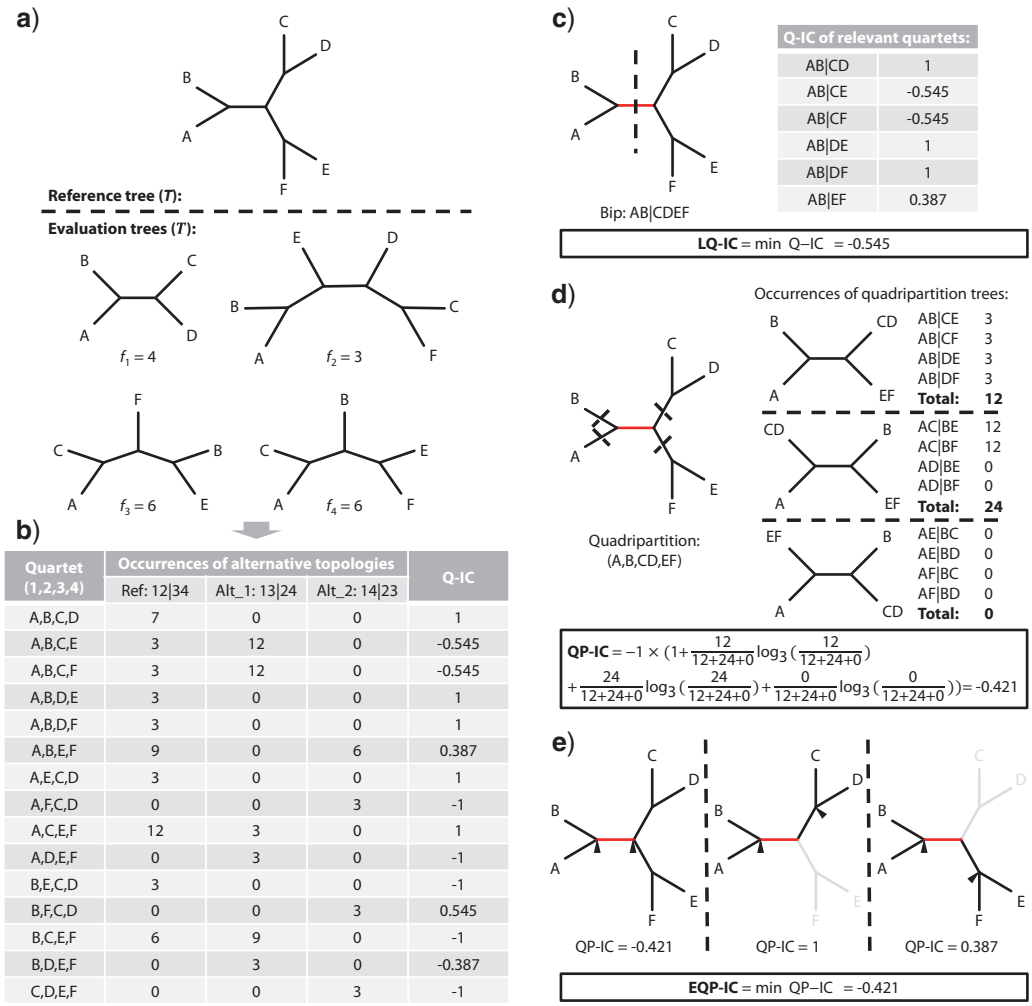
Measure 1: Lowest Quartet Internode Certainty

We define the lowest quartet internode certainty (LQ-IC) of an internal branch as the lowest IC score among all of its relevant quartets (Fig. 1b). In brief, in a given unrooted tree, every internal branch defines a nontrivial bipartition, that is, it divides the taxon set into two nontrivial subsets of taxa. We say a quartet q is relevant to an internal branch i if q consists of exactly two taxa from each of the two taxon subsets associated with i. For each internal branch i in T_R, we first identify the collection of all quartets (Q) that are relevant to i, and then calculate the IC score for each quartet q in Q based on the occurrences of its three possible topologies in T_E (c_1, c_2, and c_3 for the reference topology q_1 and the two alternative topologies, q_2 and q_3, respectively):

$$\text{Q-IC (Quartet-IC) score} = \begin{cases} 0, if\ P(q_1)=P(q_2)=P(q_3)=0; \\ 1+P(q_1)\log_3(P(q_1))+P(q_2)\log_3(P(q_2)) \\ \quad +P(q_3)\log_3(P(q_3)), \\ if\ P(q_1)\geq P(q_2)\ \text{and}\ P(q_1)\geq P(q_3); \\ -1*(1+P(q_1)\log_3(P(q_1))+P(q_2)\log_3(P(q_2)) \\ \quad +P(q_3)\log_3(P(q_3))), else. \end{cases} \quad (1)$$

where $P(q_1)=c_1/(c_1+c_2+c_3)$, $P(q_2)=c_2/(c_1+c_2+c_3)$, $P(q_3)=c_3/(c_1+c_2+c_3)$. In cases where the topology q_i has an occurrence of 0 (i.e., $P(q_i)=0$), we follow the convention that $P(q_i)\log_3(P(q_i))=0$. The Q-IC score is defined to be 0 if q does not appear in any evaluation tree (i.e., $c_1=c_2=c_3=0$). Also, we reverse the sign of the score if the topology of q induced by T_R is less frequent than any of the two alternative topologies. It should be noted that this negation is merely to indicate that the most prevalent topology conflicts with T_R.

Similar to IC/ICA scores, the Q-IC score can take values between –1 and 1: it approaches 1 when the reference quartet tree topology is much more prevalent than the other two alternatives, reflecting strong confidence in the reference internal branch; it becomes close to 0 when the three alternative topologies have similar frequencies, suggesting a high level of incongruence; and it gets near –1 when one of the conflicting topologies has a much higher frequency than the reference internal branch, indicating that the evaluation trees strongly contradict the internal branch present in the reference topology and favor an alternative topology. A visualization of the Q-IC score against possible combinations of $P(q_1)$, $P(q_2)$, and $P(q_3)$ values is provided in Supplementary Figure S1 available on Dryad at http://dx.doi.org/10.5061/dryad.440874g.



Quartet (1,2,3,4)	Occurrences of alternative topologies			Q-IC
	Ref: 12|34	Alt_1: 13|24	Alt_2: 14|23	
A,B,C,D	7	0	0	1
A,B,C,E	3	12	0	-0.545
A,B,C,F	3	12	0	-0.545
A,B,D,E	3	0	0	1
A,B,D,F	3	0	0	1
A,B,E,F	9	0	6	0.387
A,E,C,D	3	0	0	1
A,F,C,D	0	0	3	-1
A,C,E,F	12	3	0	1
A,D,E,F	0	3	0	-1
B,E,C,D	3	0	0	-1
B,F,C,D	0	0	3	0.545
B,C,E,F	6	9	0	-1
B,D,E,F	0	3	0	-0.387
C,D,E,F	0	0	3	-1

Q-IC of relevant quartets:	
AB|CD	1
AB|CE	-0.545
AB|CF	-0.545
AB|DE	1
AB|DF	1
AB|EF	0.387

AB|CE	3
AB|CF	3
AB|DE	3
AB|DF	3
Total:	**12**
AC|BE	12
AC|BF	12
AD|BE	0
AD|BF	0
Total:	**24**
AE|BC	0
AE|BD	0
AF|BC	0
AF|BD	0
Total:	**0**

FIGURE 1. An example data set to illustrate the design and calculation of quartet-based IC scores. a) The data set consists of a six-species reference tree T_R and an evaluation tree setT_E, each with a given frequency (shown along the respective tree topology). b) The reference tree is decomposed into 15 quartets and the occurrences of their quartet tree topologies in the evaluation tree set are counted. This example focuses on the internal branch that separates (A, B) from (C, D, E, F). c) The LQ-IC (lowest quartet IC) score of the internal branch is defined as the lowest IC score among all of its relevant quartets. d) The QP-IC (quadripartition IC) score of the internal branch is defined as the IC score of the quadripartition induced by it. e) The EQP-IC (extended quadripartition IC) score of an internal branch is defined as the lowest QP-IC score among all of its relevant internal node pairs.

To obtain the LQ-IC score, we simply assign the lowest Q-IC score from Q to i:

$$\text{LQ-IC (Lowest Quartet-IC) score} = \min_{q \in Q} (\text{Q-IC}(q)) \quad (2)$$

A detailed rationale on why we chose the lowest (minimum) Q-IC score, instead of other statistics such as the mean or the median is provided in the Supplementary Text available on Dryad. Since the calculation of the LQ-IC score for a given internal branch does not make any assumption about the topology on either side of i, LQ-IC can also be calculated for a reference tree T_R that contains polytomies (multifurcations); trees in the set of evaluation trees T_E may also contain polytomies. Unresolved quartets caused by polytomies in either T_R or T_E would be ignored and thus do not contribute to the calculation of the LQ-IC score.

To analyze the example data set shown in Figure 1, we first decompose the reference tree into 15 quartets and, for each quartet, we calculate the occurrences of the three possible topologies in the evaluation tree set (Fig. 1b). The IC score of each quartet can be determined from the occurrences of its alternative topologies by equation (1) (Fig. 1b). For instance, for the quartet (A, B, C, E), the reference topology (AB|CE) and the two alternative topologies (AC|BE) and (AE|BC) are respectively observed 3, 12, and 0 times in the evaluation trees. Thus, for this quartet:

$$\text{Q-IC} = -1 * \left(\frac{3}{3+12+0} \log_3 \left(\frac{3}{3+12+0} \right) + \frac{12}{3+12+0} \log_3 \left(\frac{12}{3+12+0} \right) + \frac{0}{3+12+0} \log_3 \left(\frac{0}{3+12+0} \right) + 1 \right) = -0.545$$

Note that the Q-IC score is negative since the reference quartet tree topology is less frequent than one of the conflicting topologies. Six of the 15 quartets are relevant to the internal branch of interest; therefore, the lowest Q-IC score among them (-0.545) is the LQ-IC score of the internal branch (Fig. 1c).

Measure 2: Quadripartition Internode Certainty

We define the quadripartition internode certainty (QP-IC) of an internal branch as the IC score of its induced quadripartition (Fig. 1c). In a given unrooted binary tree, each internal branch connects two internal nodes (hereafter referred to as nodes) and divides the taxon set into four subsets (quadripartition). To determine the IC score of an internal branch, we assume that the four subsets have been correctly resolved and only consider the three possible topologies of the quadripartition. In other words, we consider the quadripartition as a "meta-quartet" whose leaves are the four subsets, and use the IC score of the "meta-quartet" tree as that of the internal branch.

For the quadripartition p induced by a given internal branch i in T_R, we calculate its IC score based on the occurrences of its three possible topologies in T_E (c_1, c_2, and c_3 for the alternative topologies p_1, p_2, and p_3, respectively). We first identify the collection of all quartets (Q) that are relevant to p; we say a quartet q is relevant to a quadripartition p if q consists of exactly one taxon from each of the four-taxon subsets associated with p. Each given quadripartition tree topology t_p, induces a specific quartet tree topology t_q for each q in Q, and the occurrence of t_p is simply the sum of the occurrence of t_q for all q in Q. We can then calculate the quadripartition IC score as:

$$\text{QP-IC (Quadripartition-IC) score} = \begin{cases} 0, \textit{if } P(p_1) = P(p_2) = P(p_3) = 0; \\ 1 + P(p_1)\log_3(P(p_1)) + P(p_2)\log_3(P(p_2)) \\ \quad + P(p_3)\log_3(P(p_3)), \\ \textit{if } P(p_1) \geq P(p_2) \text{ and } P(p_1) \geq P(p_3); \\ -1 * \Big(1 + P(p_1)\log_3(P(p_1)) + P(p_2)\log_3(P(p_2)) \\ \quad + P(p_3)\log_3(P(p_3)) \Big), \textit{else}; \end{cases} \tag{3}$$

where $P(p_1) = c_1/(c_1+c_2+c_3)$, $P(p_2) = c_2/(c_1+c_2+c_3)$, and $P(p_3) = c_3/(c_1+c_2+c_3)$. Note that the equations for the calculation of Q-IC (1) and QP-IC (3) are almost the same, except that the Q-IC score is calculated from a single quartet, whereas the QP-IC score from a quadripartition. If the quadripartition tree topology induced by T_R is less frequent than any other alternative topologies, we reverse the sign of the QP-IC score. Unlike the LQ-IC score, the QP-IC is only calculated for internal branches in the reference tree T_R that are fully resolved on both sides (QP-IC scores are left undefined for polytomies). The evaluation trees, however, may contain polytomies.

In the example data set, the quadripartition induced by the internal branch of interest is (A, B, CD, EF). The occurrence of each alternative quadripartition tree topology equals the sum of the occurrences of its induced quartet tree topologies (Fig. 1d). For instance, the reference quadripartition, (A, B|CD, EF), induces four quartet trees: (AB|CE), (AB|CF), (AB|DE), and (AB|DF). Each quartet tree is observed three times in the evaluation trees. Therefore, the quadripartition tree has a total occurrence of 12. In the same way, the occurrences of the two conflicting topologies, (A, CD|B, EF) and (A, EF|B, CD), are 24 and 0, respectively. Thus, for this quadripartition:

$$\text{QP-IC} = -1 * \left(\frac{12}{12+24+0} \log_3 \left(\frac{12}{12+24+0} \right) + \frac{24}{12+24+0} \log_3 \left(\frac{24}{12+24+0} \right) + \frac{0}{12+24+0} \log_3 \left(\frac{0}{12+24+0} \right) + 1 \right) = -0.421$$

Once again, the QP-IC score is negative since the reference quadripartition is less frequent than at least one of the conflicting topologies.

Measure 3: Extended Quadripartition internode certainty

In measure 2, we only consider quadripartitions induced by individual internal branches, which means that only the quartets corresponding to neighboring nodes contribute to the IC score calculation. In particular, quartets that contradict an internal branch but are not relevant to the corresponding quadripartition are ignored by the QP-IC score. An alternative approach is to extend measure 2 to evaluate all possible pairs of nodes that include a given internal branch (see Supplementary Text available on Dryad for an extended discussion of the rationale of extended quadripartition internode certainty, EQP-IC). The design of EQP-IC is similar to that of LQ-IC, with a critical distinction as, here, we examine the IC scores of node pairs instead of individual quartets. We define the EQP-IC score of an internal branch as the lowest IC score among all of its relevant node pairs (N) (Fig. 1e); we say a pair of nodes $n_{(i)}$ is relevant to

an internal branch i if i is part of the path connecting $n_{(i)}$ (note that there is no upper limit on the length of the path). Apparently, $N_{(i)}$ includes both neighboring and nonneighboring node pairs. The IC score of a pair of neighboring nodes is simply its QP-IC score (see measure 2). In a given unrooted binary tree, every node has three outgoing branches. We can therefore construct a "meta-quartet" from the path connecting a pair of nonneighboring nodes as well as the two outgoing branches of each node that are not located on the path, and determine its QP-IC score also using measure 2. To obtain the EQP-IC score, we simply assign the lowest QP-IC score from $N_{(i)}$ to internal branch i:

$$\text{EQP-IC (Extended Quadripartition-IC) score} = \min_{n_{(i)} \in N_{(i)}} (\text{QP-IC } (n_{(i)})) \quad (4)$$

To calculate the EQP-IC score, the reference tree T_R must be binary, but the evaluation trees may be polytomous.

In the example data set, the reference tree contains three node pairs that are relevant to the internal branch of interest (Fig. 1e). The first one is the neighboring node pair that defines the internal branch itself. Hence, its QP-IC score equals that of the internal branch. The other two nonneighboring node pairs induce the quadripartitions (A, B, C, D) and (A, B, E, F), respectively, and their QP-IC scores are found to be 1 and 0.387 by applying the same procedure as described in the preceding section. Consequently, the lowest QP-IC score among the three node pairs (-0.421) is assigned to be the EQP-IC score of the internal branch.

Results and Discussion

Quartet-Based and Bipartition-Based Measures Yield Similar IC Scores on Complete Trees

We compared the performances of the quartet-based and bipartition-based measures on a simulated data set consisting of 50 reference trees, each comprising 101 taxa; each reference tree is then associated with 1000 complete evaluation trees (hereafter referred to as the "G1000_Original" data set; see Materials and Methods section). The relative Robinson–Foulds (rRF) distance between the reference and evaluation trees ranges from 0.19 to 1 with a median value of 0.43. The three quartet-based IC scores are almost perfectly correlated with each other and the same is true for the bipartition-based IC/ICA scores (Spearman correlation coefficients ≥ 0.99 and P-values $< 2.2 \times 10^{-16}$ in all cases; Supplementary Fig. S2 available on Dryad). Moreover, quartet-based IC scores are strongly correlated with branch support values (measured by Gene Support Frequency; Gadagkar et al. 2005) and bipartition-based IC/ICA scores (Spearman correlation coefficients ≥ 0.93 and P-values $< 2.2 \times 10^{-16}$ in all cases; Supplementary Fig. S2 available on Dryad).

The above analysis uses species trees as reference trees and gene trees as evaluation trees. However, this is not a requirement (Salichos et al. 2014). For example, one could use gene trees as reference trees and their corresponding bootstrap replicate trees as evaluation trees. Therefore, we also analyzed such a data set that contained 844 maximum-likelihood gene trees (used as reference trees), each associated with 200 bootstrap replicate trees (used as evaluation trees) (hereafter referred to as the "B200" data set). The quartet-based IC scores again show strong correlation with bootstrap supports and bipartition-based IC/ICA scores (Spearman correlation coefficients ≥ 0.91 and P-values $< 2.2 \times 10^{-16}$ in all cases; Supplementary Fig. S3 available on Dryad). Overall, our results suggest that the IC scores generated by quartet-based and bipartition-based measures are generally in agreement on data sets that only comprise complete trees.

Quartet-Based IC Measures are More Robust in Data Sets with Missing Data

Next, we compared the performance of quartet-based and bipartition-based IC measures on data sets with missing data. To that end, we constructed a series of additional data sets with varying degrees of missing data at the taxon and gene level (see Materials and Methods section for details). In brief, we first generated five data sets with partial trees—named G1000_L1/L2/L3/E1/E2—by pruning taxa from evaluation trees in the G1000_Original data set. We note that: (1) the pruned taxa were randomly selected in G1000_L1/L2/L3, while the patterns of missing taxa in G1000_E1/E2 were sampled from empirical data sets and (2) the degree of missing taxa increases in sequential order in G1000_L1/L2/L3, while the degrees of missing taxa in G1000_E1 and G1000_E2 are comparable to those in G1000_L2 and G1000_L3, respectively (Supplementary Fig. S4 available on Dryad). Additionally, we randomly removed evaluation trees from the G1000_Original/L1/L2/L3/E1/E2 data sets to create two collections of data sets with removed genes, G500 and G250, in which each reference tree is associated with 500 and 250 evaluation trees, respectively.

To examine the performance of each measure, we followed Kobert et al. (2016) suggestion that, on data sets with missing data, a more robust measure should give scores that are closer to the ground truth. Here, we measured the robustness to missing data by the Euclidean distance between the IC scores calculated on the G1000_Original data set (which we consider as the "truth") and the pruned data sets. A smaller Euclidean distance indicates higher robustness, and vice versa. The Euclidean distances were calculated for each of the 50 sets of reference/evaluation trees in each pruned data set, and for each of the quartet-based and bipartition-based measures. For ease of comparison, all Euclidean distances were normalized to the same scale between 0 and 1 (see Material and Methods section).

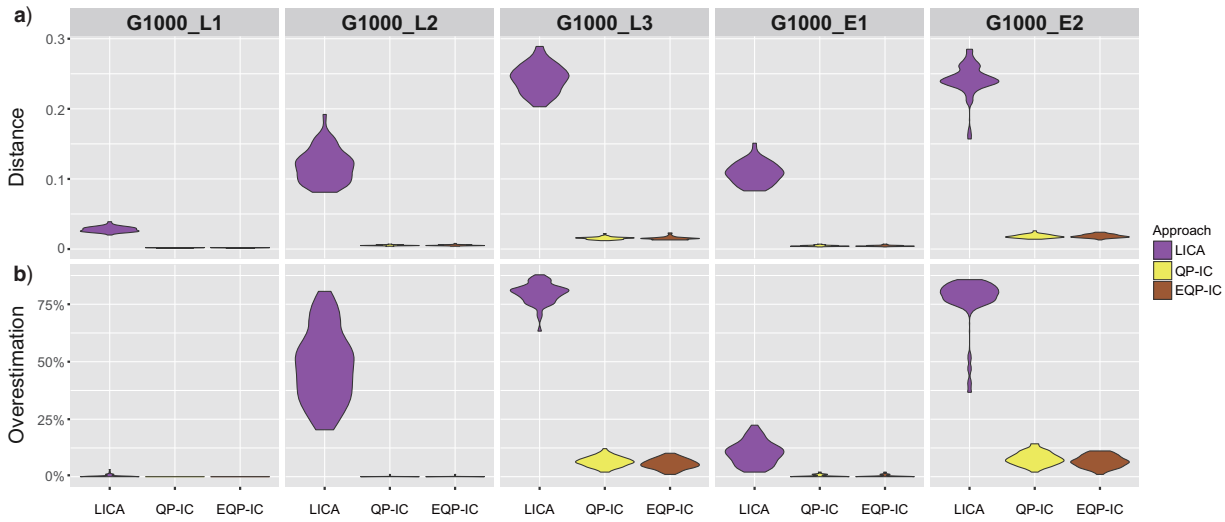


FIGURE 2. Quartet-based IC measures are more robust on partial evaluation trees. a) Euclidean distances between IC scores calculated from the G1000_Original data set, which contain only complete evaluation trees, and those calculated from data sets G1000_L1/L2/L3 and G1000_E1/E2, which contain partial evaluation trees. b) Fractions of internal branches for which the IC scores were overestimated (0.05 unit higher) on data sets G1000_L1/L2/L3 and G1000_E1/E2 compared with the G1000_Original data set. The violin plots depict, for each measure, distribution of Euclidean distance (a) or overestimated fraction (b) values from 50 replicates.

The results show that QP-IC and EQP-IC exhibit high robustness (median distances of 0.04 or less) on all data sets examined (Fig. 2 and Supplementary Fig. S5 available on Dryad), whereas the robustness of LQ-IC is much lower (Supplementary Fig. S5 available on Dryad). Thus, we have focused to present and discuss the results of QP-IC and EQP-IC; for Results and Discussion section about LQ-IC, we refer the reader to the Supplementary Text available on Dryad. On the other hand, the robustness of bipartition-based IC measures is highly dependent on the proportion of missing taxa. For instance, the median distances for lossless internode certainty (LIC)/LIC (all) (LICA) are less than 0.04 on the Original and L1 data sets, but increase to > 0.10 and > 0.20 on the L2/E1 and L3/E2 data sets, respectively (Fig. 2 and Supplementary Fig. S5 available on Dryad; for simplicity, only ICA and/or LICA are shown in the main figures as the representatives of bipartition-based IC measures, while the full results are included in Supplementary figures available on Dryad).

For each measure, we also calculated the fractions of internal branches for which the scores were overestimated on the pruned data sets compared with the unpruned data set. We observe a trend very similar to that found in the robustness assessment. Quartet-based measures exhibit significantly lower levels of overestimation than bipartition-based measures on data sets with medium to high proportions of missing taxa (paired Wilcoxon rank-sum test, P-values $< 2.2 \times 10^{-16}$ for all pairwise comparisons between quartet-based and bipartition-based IC measures on each L2/L3/E1/E2 data sets; Fig. 2 and Supplementary Fig. S5 available on Dryad). In particular, the LIC and LICA scores are consistently overestimated at high proportions of missing taxa (Fig. 2 and Supplementary Fig. S5 available on Dryad); the scores are overestimated for more than 75% of all internal branches in the L3 and E2 data sets, whereas the overall fractions of overestimated quartet-based IC scores on the same data sets are below 30%. Altogether, these results suggest that quartet-based IC measures are more robust than bipartition-based measures on partial evaluation trees.

We further compared the quartet-based and bipartition-based IC measures on two empirical data sets previously analyzed in Kobert et al. (2016), namely a 23-taxon yeast data set containing 1275 complete gene trees and 1219 partial gene trees, and an avian data set containing 500 complete gene trees and 1500 partial gene trees. IC scores calculated from all gene trees were compared with the scores calculated from either only the complete trees or only the partial trees. Here again, we found that the Euclidean distances for quartet-based measures are lower than the distances for bipartition-based measures in all comparisons (Supplementary Fig. S6 available on Dryad). Particularly, in the comparison between scores calculated from either all trees or partial trees only, the quartet-based measures have maximum distances of 0.04 and 0.01 on the yeast and avian data sets, respectively. In contrast, the bipartition-based measures exhibit a minimum distance of 0.12 on both data sets. The results again suggest that quartet-based IC scores are more robust in the presence of partial evaluation trees.

Importantly, our results suggest that missing genes appear to have limited impact on quartet-based IC measures. On the simulated data sets, highly similar scores can be obtained with only 25% of all evaluation trees (e.g., Euclidean distances are below 0.03 in the comparison of G1000_Original and G250_Original; Supplementary Fig. S5 available on

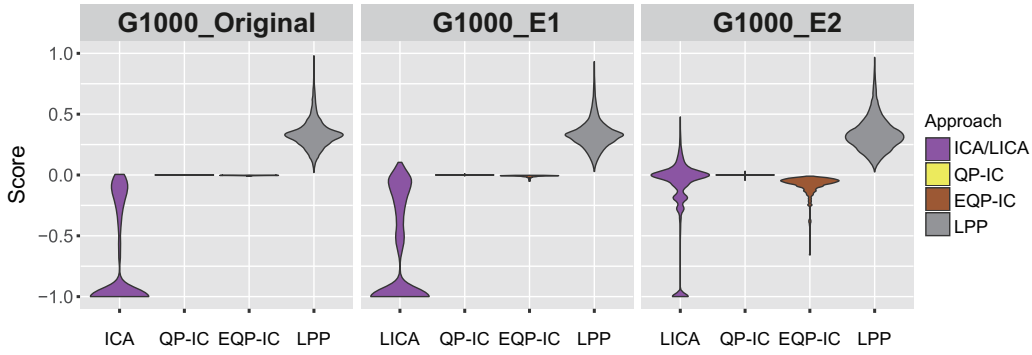


FIGURE 3. Bipartition-based IC scores tend to generate artificially low scores in the lack of phylogenetic signal. The results of the "random evaluation tree" test on data sets G1000_Original, G1000_E1, and G1000_E2 are shown. The violin plots indicate, for each measure, the distribution of IC scores or LPP supports calculated using randomized evaluation trees. It should be noted that the IC scores can take values between –1 and 1, whereas the LPP support takes values between 0 and 1.

Dryad). On the two empirical data sets, the exclusive use of either complete trees or partial trees only yields highly similar scores to analyses that include all evaluation trees (Euclidean distances are below 0.05 in all comparisons; Supplementary Fig. S6 available on Dryad). Phylogenomic data matrices can vary substantially with respect to the number of genes used and amount of missing data. Therefore, the observation that quartet-based IC measures are robust to missing genes suggests that these measures can provide robust estimates of phylogenetic incongruence from incomplete phylogenomic data sets.

Additionally, this property is particularly desirable for analyzing bootstrap trees as the number of bootstrap replicates is typically rather small (e.g., 100). We therefore randomly sampled 100 and 50 bootstrap replicate trees per gene tree from the B200 data set and examined the robustness of the IC measures. The results show that, indeed, the Euclidean distances remain very low for all IC measures (median distances less than 0.05) even when only 50 bootstrap trees are used in the evaluation (Supplementary Fig. S7 available on Dryad).

Robustness and Limitations of IC Measures to Specific Analytical Challenges

To investigate the potential strengths and/or weaknesses of different IC measures, we next assessed their performance under two analytical challenges, namely lack of phylogenetic signal, and topological errors in reference trees. To better assess the performance of IC scores, we also included in our comparison the local posterior probability (LPP), a coalescent-based approach to calculate branch support for species trees that also relies on quartet frequencies in gene trees (Sayyari and Mirarab 2016).

Lack of phylogenetic signal.—In the aforementioned analysis of the empirical avian data set, we observed that, for some internal branches, the quartet-based IC scores are around 0, a value indicative of two nearly equally supported conflicting resolutions, whereas the bipartition-based IC scores are near or at –1, a value indicative of the presence of a conflicting bipartition that is much more strongly supported than the reference bipartition. Closer examination of the underlying bipartition frequencies at these internal branches revealed that none of the conflicting bipartitions is supported (Supplementary Table S2 available on Dryad). For instance, for multiple internal branches, the reference bipartition and the most prevalent conflicting bipartition have frequencies of 0 and 0.03, respectively. This suggests that bipartition-based IC measures might report strong support for a conflicting bipartition when in reality there is little phylogenetic signal.

To test this behavior of bipartition-based IC scores further, we devised a "random evaluation tree" test where we used completely random evaluation tree topologies in the G1000_Original/E1/E2 data sets (see Materials and Methods section). It should be noted that all bipartitions of the same size would have equal probability to be included in a completely random tree. Thus, in principle, the evaluation tree sets should provide no support to any particular relationship and the IC scores for all internal branches should be near or at 0. The results of this test show that quartet-based measures in general are highly robust to the lack of phylogenetic signal; the scores are tightly distributed around median values that are between -0.06 and 0 (Fig. 3).

In contrast, bipartition-based IC scores (except for Probabilistic internode certainty all (PICA) scores) are heavily skewed toward –1 (Fig. 3 and Supplementary Fig. S8 available on Dryad). For instance, 64.6% of all internal branches have IC and ICA scores of –1 on the data set G1000_Original (no missing data), while 53.8% and 43.1% of all internal branches have Probabilistic internode certainty (PIC) scores of –1 on G1000_E1 (medium proportion of missing data) and G1000_E2 (high proportion of missing data), respectively (Fig. 3). Interestingly, LIC and LICA scores are at –1 for 53.0% of all internal branches on G1000_E1, but for only 10.1% on G1000_E2 (Fig. 3). On the other hand, 86.3% of PICA scores are between -0.1 and 0 on both G1000_E1 and G1000_E2 (Supplementary Fig. S8

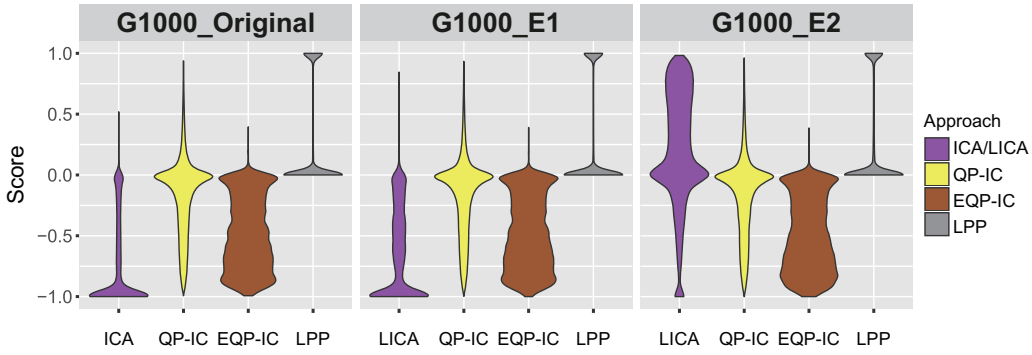


FIGURE 4. The robustness of quartet-based and bipartition-based IC measures to errors in reference trees. The results of the "altered reference tree" test on data sets G1000_Original, G1000_E1, and G1000_E2 are shown. The violin plots indicate, for each measure, the distribution of IC scores or LPP supports for bipartitions that are only present in the altered reference trees.

available on Dryad). Overall, the results suggest that bipartition-based IC measures are in general sensitive to the lack of phylogenetic signal. On all three data sets, the LPP support values appear to be distributed around ∼0.33, which corresponds to equal supports for all three alternative topologies and is thus expected given random evaluation trees (Fig. 3).

Errors in reference tree.—One important assumption underlying the design of the QP-IC and EQP-IC scores is that the four subsets of taxa around a given internal branch are correctly resolved (referred to as the "locality assumption" in Sayyari and Mirarab 2016). To test the performance of QP-IC and EQP-IC when the locality assumption is violated and also the performance of other IC measures on reference trees containing incorrect relationships, we devised an "altered reference tree" test. In this test, varying degrees of errors were introduced into the reference trees in the G1000_Original/E1/E2 data sets by replacing them with topologies selected from their respective evaluation tree sets; the corresponding rRF distances between original and altered reference trees range between 0.1 and 1 (see Materials and Methods section).

We examined the IC scores of the bipartitions that are only present in the altered reference trees but not in the original ones. Since the vast majority (91.8%; 4499 out of 4900) of internal branches in the original reference trees have positive IC scores (Supplementary Fig. S2 available on Dryad), the bipartitions introduced by the alterations are expected to be contested by other, higher-frequency bipartitions. Indeed, more than 99% of the introduced internal branches have negative EQP-IC scores as well as negative bipartition-based IC scores in the absence of missing data (data set G1000_Original; Fig. 4 and Supplementary Fig. S9 available on Dryad). Conversely, the QP-IC scores are positive for a considerable fraction (24.8%) of these introduced internal branches (Fig. 4). Interestingly, all the IC measures generate lower scores on altered reference trees that are more dissimilar to the original trees than on altered reference trees that are more similar to the original trees (Supplementary Fig. S9a available on Dryad). The same patterns

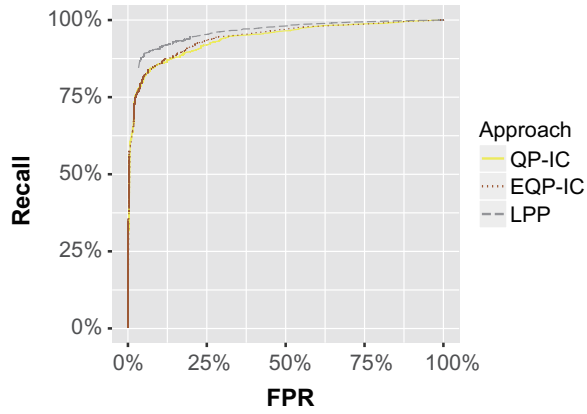


FIGURE 5. Evaluation of the performance of quartet-based IC measures in predicting branch correctness. The plots of FPR vs. recall (ROC curves) are shown for quartet-based IC measures and LPP.

are also observed on data set G1000_E1, which has a medium proportion of missing data (Fig. 4 and Supplementary Fig. S9b available on Dryad). However, at high proportion of missing data (data set G1000_E2), bipartition-based measures LIC/LICA and PIC/PICA produce positive scores for 69.9% and 32.4% of the incongruent internal branches, respectively, even greater than QP-IC (24.9%) (Fig. 4 and Supplementary Fig. S9c available on Dryad). In contrast, the performances of quartet-based IC measures are consistent at all different proportions of missing data (Fig. 4 and Supplementary Fig. S9 available on Dryad).

These results suggest that the violation of the locality assumption can often lead to inflated QP-IC scores. LPP, which also relies on the locality assumption, yields support values greater than 0.95 for 20.1%∼22.5% of the introduced internal branches in each of the three data sets (Fig. 4). The related EQP-IC measure is more robust to such violations; the locality assumption might be relaxed due to the consideration of other nonneighboring node pairs in the EQP-IC measure. Bipartition-based IC measures perform generally well except at high proportion of missing data. The reason might be that, as has been shown earlier in this study, bipartition-based measures tend to overestimate IC

scores when the proportion of missing data is high (Fig. 2).

Related Measures of Branch Support or Phylogenetic Incongruence

Besides the IC scores discussed above, there also exist other related measures of branch support or phylogenetic incongruence. The LPP is a fast method for local branch support (Sayyari and Mirarab 2016). Similar to QP-IC, it first calculates the quartet-based frequencies of alternative quadripartition topologies around a given internal branch in the species tree from a set of gene trees. The probability that the quadripartition is present in the true species tree (i.e., the LPP), is then estimated under the multi-species coalescent (MSC) model. By invoking the MSC model, LPP explicitly accounts for ILS which is an important source of species tree-gene tree discordance. LPP has been shown to exhibit high precision and sensitivity on data sets simulated under the MSC model (Sayyari and Mirarab 2016). However, the performance of LPP might be compromised if the incongruence is driven by processes other than ILS. In comparison, neither our quartet-based nor the original bipartition-based IC measures make any assumption on the underlying causes of incongruence. Therefore, these measures are broadly applicable to measuring the level of incongruence in any data type (e.g., a maximum-likelihood gene tree as the reference tree and the corresponding bootstrap replicate trees as evaluation trees).

An important difference between IC measures and branch support measures, such as LPP, is their behavior with respect to the amount of data available. Increased numbers of gene trees should in principle increase our confidence in the species tree inference and will thus likely increase LPP support values; however, the IC measures quantify the degree of conflict and their scores will therefore not necessarily change. Consider a data set with a species tree and 50 gene trees where the only source of species tree-gene tree incongruence is ILS. Suppose that there is an internal branch in the species tree for which the three possible quadripartitions have frequencies of 40%, 30%, and 30%, respectively. The LPP support and QP-IC score for this internal branch will be 0.65 and 0.01, respectively. Now consider how the LPP support value and the QP-IC score change if the number of gene trees is increased, while the frequencies of quadripartitions remain unaltered. With 200, 500, and 2000 gene trees, the LPP supports will increase to 0.95, 0.99, and 1, respectively (see Sayyari and Mirarab 2016), whereas the QP-IC score remains at 0.01 because the degree of incongruence is constant. This increase in LPP support is expected under the MSC model (in which the LPP is determined by both the frequencies of alternative quadripartition topologies and the total number of genes), but might be problematic if species tree-gene tree incongruence is predominantly caused by processes other than ILS.

It is important to note that all IC measures have been specifically designed to quantify the level of incongruence in phylogenetic data sets, whereas branch support measures, such as bootstrap and posterior probability, aim to predict the correctness of the tree topology. However, the IC scores might still be correlated with branch correctness given that QP-IC and LPP are calculated from the same underlying data (i.e., frequencies of the three possible topologies of each quadripartition). We therefore also evaluated the performance of quartet-based IC scores in predicting branch correctness. We used the G1000_E2 data set in which both estimated and true species trees are available to compare the performance of quartet-based IC scores and LPP support values in terms of false positive rate (FPR; the percentage of false branches that are predicted to be correct), recall (the percentage of true branches that are predicted to be correct), and precision (the percentage of true branches among the ones predicted to be correct). Ideally, branch support methods should have low FPR, high recall, and high precision.

Our results show that, for QP-IC, the FPR is 100% at a threshold of 0 but quickly drops to 1.9% at a threshold of 0.1, whereas the precision increases from 94.5%—which equals the overall percentage of true branches—to 99.9% (Supplementary Fig. S10 available on Dryad; note that both FPR and precision remain almost constant at thresholds below 0 or above 0.1). In other words, when we use a QP-IC score of 0.1 as a threshold, only 1.9% of all false branches are predicted to be correct, and 94.5% of the branches predicted to be correct are true. At the same time, the recall gradually decreases from 100% to 69.6% (Supplementary Fig. S10 available on Dryad). EQP-IC exhibits an almost identical behavior (Supplementary Fig. S10 available on Dryad). In comparison, LPP has a higher FPR (9.7% and 3.0% at thresholds of 0.99 and 1, respectively), better recall (90.6% and 76.5%, respectively), and similar precision (99.4% and 99.8%, respectively). We further used receiver operating characteristics (ROC) curves to compare the quartet-based IC measures and the LPP for their performance as diagnostic tests. The plots in Figure 5 show that LPP achieves the best tradeoff between FPR and recall, and QP-IC/EQP-IC are closely behind. Overall, although QP-IC and EQP-IC have not been designed as branch support measures, our results suggest that they perform well in predicting branch correctness. It should be noted, however, that the evaluation here is based on simulated data sets in which the only sources of species tree-gene tree discordance are ILS and species/gene tree estimation error.

In parallel to this work, Pease et al. (2018) developed the quartet sampling (QS) measure, which is also a quartet- and entropy-based measure of phylogenetic incongruence (like IC measures). The major distinction between QS and our QP-IC measure is that, in QS, quartets are randomly sampled and the three alternative topologies for each quartet are evaluated independently under the ML criterion, whereas in QP-IC, all quartet

tree topologies are extracted from already estimated evaluation trees. Accordingly, QS requires only the reference tree but can only be applied to a single data matrix; on the other hand, our quartet-based measures require pre-estimated evaluation trees but can be used for both single data matrix analysis (on bootstrap replicate trees) and for coalescent analysis (on single-gene trees).

Additional studies are needed to compare the performances of QS and our quartet-based IC measures on phylogenomic data sets. On one hand, the evaluation of quartet tree topologies in the QS measure might be sensitive to phylogenetic artifacts such as long-branch attraction (Ranwez and Gascuel 2001). On the other hand, the performance of quartet-based IC measures can be impaired by inaccurate gene tree estimation when the numbers of taxa become high and the lengths of single-gene alignments become short. Nevertheless, the two types of measures can complement each other and their joint usage in phylogenomic studies will likely yield a more comprehensive understanding of phylogenetic incongruence in genome-scale data sets.

Finally, the idea of QS is also worth future exploration. Currently, each evaluation tree contributes $\binom{n}{4}$ (n equals the number of taxa in the evaluation tree) quartets to the calculation of quartet-based IC scores, and thus the evaluation trees with missing taxa might arguably be over-penalized. In addition, the major computational bottleneck in the calculation of LQ-IC and EQP-IC scores is to count the occurrences of all quartet tree topologies in the evaluation tree set. QS might provide a viable solution for both limitations.

Analysis of an Empirical Phylogenomic Data Set

We lastly applied our quartet-based IC measures to a data set of 42 therian mammals and 5246 genes that was recently used by Scornavacca and Galtier (2017) to investigate the role of ILS in mammalian phylogenomics. The evolutionary history of mammals has been difficult to resolve and is one of the most debated questions in phylogenetics (Foley et al. 2016). The phylogeny remains partly unsettled despite of considerable efforts over the years, as well as recent progress in data acquisition and analytical techniques (e.g., Song et al. 2012; Tarver et al. 2016; Esselstyn et al. 2017; Scornavacca and Galtier 2017).

As shown in Figure 6, internal branches in the mammalian phylogeny display substantially different degrees of phylogenetic incongruence. For instance, their EQP-IC scores range from 0 to 0.99, with a median of 0.26; 12 of the 39 internal branches have EQP-IC scores exceeding 0.5, indicating that the frequency of the reference topology must be at least 76%, whereas 11 other internal branches have scores below 0.1, corresponding to reference topology frequencies of 57% or less. The highest levels of incongruence are observed for the most controversial relationships in mammalian phylogeny, including: (1) the root of placental mammals (score of 0.02); (2) the placement of tree shrew within Euarchontoglires (score of 0); (3) the early divergence of rodents (score of 0); and (4) the diversifications at the base of Laurasiatheria (scores between 0 and 0.04) (Fig. 6). QP-IC produced highly similar scores to those of EQP-IC (Spearman correlation coefficient ≥ 0.99 and P-value $< 2.2 \times 10^{-16}$; Fig. 6).

In sharp contrast, all but one of the internal branches have LPP support values of 1 (Fig. 6). LPP is built on the MSC model and effectively attributes all observed incongruence to ILS. Therefore, high LPP support values for internal branches with low IC scores would be reasonable if the incongruence is largely due to ILS; however, if the underlying cause(s) of incongruence are different, then the high LPP support values may be misleading. Indeed, several recent studies have provided evidence that the majority of phylogenetic conflicts in mammals are likely due to factors other than ILS (Tarver et al. 2016; Esselstyn et al. 2017; Scornavacca and Galtier 2017). Overall, this empirical example demonstrates that our quartet-based IC measures represent useful diagnostic tools for identifying potentially problematic phylogenetic relationships.

Implementation

We have implemented the three quartet-based IC measures in the program *QuartetScores*, which is freely available as open source code at https://github.com/lutteropp/QuartetScores. Several existing algorithms (e.g., ASTRAL, Mirarab et al. 2014; tqDist, Sand et al. 2014) can efficiently determine the total number of quartets shared between trees and can thus be readily used for calculating QP-IC scores. However, the calculation of LQ-IC and EQP-IC scores still requires counting the occurrences of individual quartet topologies which represents a major computational bottleneck. Therefore, we devised two algorithms for quartet counting: one is more time-efficient by storing each quartet topology separately in a lookup table; the other is more memory-efficient by grouping different topologies of a quartet together and using a more involved indexing function. The program will automatically decide which algorithm to use based on the data set size. A full description of the algorithms for counting quartets and computing quartet-based IC scores is provided in the Supplementary Text available on Dryad.

Concluding Remarks

In this study, we have introduced three quartet-based IC measures, namely LQ-IC, QP-IC, and EQP-IC. Much like previously bipartition-based IC measures (Salichos et al. 2014; Kobert et al. 2016), these new quartet-based IC measures are information theory-based measures to quantify phylogenetic incongruence and are calculated from the frequencies of each of the three possible quartet tree topologies using Shannon's entropy function. They

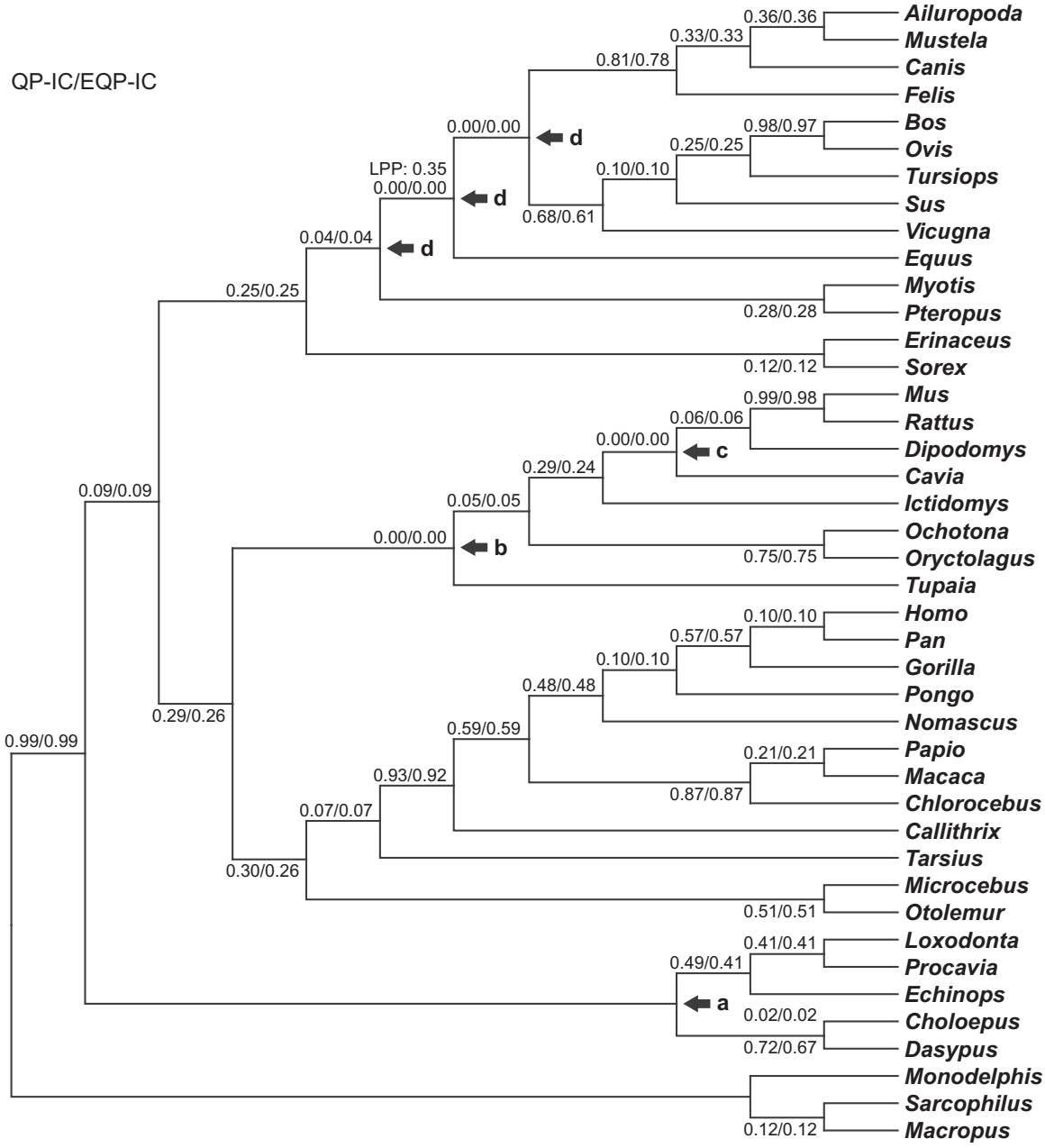


FIGURE 6. Analysis of the empirical mammalian phylogenomic data set using quartet-based and bipartition-based IC measures. Values shown at each internal branch indicate QP-IC and EQP-IC scores. All but one of the internal branches have LPP support of 1. The labeled internal branches correspond to (a) the root of placental mammals; (b) the placement of tree shrew within Euarchontoglires; (c) the early divergence of rodents; and (d) the diversifications at the base of Laurasiatheria.

summarize the diversity and strength of conflicting signals via a single number; score values close to 1 (or −1) suggest that the given internal branch is supported (or contested), whereas score values close to 0 indicate high levels of incongruence or the lack of phylogenetic signal. Each specific value of the IC score corresponds to a range of possible combinations of frequencies for the three alternative quartet tree topologies. The decision of whether a specific IC score is high or low is nonetheless subjective. To facilitate the interpretation of IC scores, we have provided a plot (Supplementary Fig. S1 available on Dryad) to display the IC scores for various combinations of frequencies of alternative quartet tree topologies.

By analyzing both simulated and empirical data sets, we have carefully compared the performance of all IC measures under various data characteristics (see Table 2 for a summary). The results show that the new quartet-based IC measures clearly outperform previous bipartition-based IC measures, and should thus be

TABLE 2. Performance of quartet-based IC measures under different analytical challenges

Measure		Definition	Performance			
			Missing gene	Missing taxa	Lack of phylogenetic signal	Errors in reference tree
Bipartition-based	IC/ICA	Shannon's entropy calculated from the frequencies of the reference and conflicting bipartitions	Robust	Sensitive	Sensitive	Robust
	PIC/PICA/ LIC/LICA	Extension of IC/ICA to the analysis of incomplete evaluation trees	Robust	Sensitive	Sensitive	Sensitive at high proportions of missing taxa
Quartet-based	LQ-IC	Lowest IC score among all relevant quartets of an internal branch	Robust	Sensitive at high proportions of missing taxa	Sensitive at high proportions of missing taxa	Robust
	QP-IC	IC score of the quadripartition around an internal branch	Robust	Robust	Robust	Sensitive
	EQP-IC	Lowest QP-IC score among all relevant node pairs of an internal branch	Robust	Robust	Robust	Robust

preferred in future phylogenomic studies. Among our quartet-based IC measures, EQP-IC is robust under all analytical challenges we have examined and is thus our recommended approach. QP-IC has a similar performance to EQP-IC, with the exception that it may assign inflated IC scores to incorrect relationships in the reference tree. For LQ-IC, a critical limitation is that it can be driven by quartets that are infrequent but have low Q-IC scores by chance. Therefore, it should be used with caution on data sets with high proportions of missing taxa.

MATERIALS AND METHODS

Simulated Data Sets

We used a simulated data set from Mirarab and Warnow (2015) (referred to as the "G1000_Original" data set in our study) to evaluate the performance of our quartet-based IC scores. This G1000_Original data set contains 50 sets of trees, each of which has one species tree and 1000 estimated gene trees. In brief, for each set, a 101-taxon species tree was first simulated according to the Yule process with a speciation rate of 10^{-6} per generation and a tree length of 2 million generations. Then, 1000 gene trees were simulated on the species tree under the multiple-species coalescent model. For each simulated gene tree, a gap-free nucleotide alignment was simulated under the GTR+GAMMA model, and FastTree 2 was used to infer a maximum-likelihood gene tree. The simulated species trees and their corresponding estimated gene trees were downloaded from https://goo.gl/KhuQtq and used in this study.

All gene trees in the G1000_Original data set are complete. To further examine the performance of quartet-based IC scores on partial trees, we generated five additional data sets by pruning taxa from gene trees in the G1000_Original data set. The taxon-pruning was conducted in two different ways. For three of the five data sets (G1000_L1, G1000_L2, and G1000_L3), taxa were pruned randomly and, for each gene tree, the number of taxa to prune was drawn from a log-normal distribution (truncated on the right at 97 to ensure that pruned trees have at least four taxa). The three data sets were generated by using log-normal distributions with mean values of ln 1, ln 10, and ln 100, respectively, corresponding to low, medium, and high proportions of missing data.

For the other two data sets (G1000_E1 and G1000_E2), the patterns of missing taxa were sampled from empirical data sets to better approximate real data conditions. For instance, G1000_E1 was generated by using the 144-taxon, 1478-gene empirical data set from Misof et al. (2014) as template as follows: first, 101 taxa and 1000 gene trees were randomly selected from the empirical data set and randomly paired with the taxa and gene trees in the G1000_Original data set; then, for each taxon t and gene tree g selected from the empirical data set, if t is missing from g, the paired taxon t' was pruned from the paired gene tree g' in the simulated G1000_Original data set. This procedure was performed independently for each of the 50 sets of trees in the G1000_Original data set. The data set G1000_E2 was constructed in the same way based on the 103-taxon, 844-gene empirical data set from Wickett et al. (2014) (note that some genes were sampled twice from the empirical data set, since it has less than 1000 genes).

To further examine the impact of missing data on IC scores, we pruned gene trees from the six above-mentioned data sets (i.e., G1000_Original/L1/L2/L3/E1/E2) to generate two additional data sets, namely G500 and G250, standing for 500 and 250 gene trees per species tree, respectively. The gene-tree pruning was conducted by randomly selecting 500 (for G500 data sets) or 250 (for G250 data sets) gene trees from each of the 50 sets of trees in the

G1000_Original data set, and the corresponding, taxon-pruned gene trees in the G1000_L1/L2/L3/E1/E2 data sets were selected accordingly.

All reference and evaluation trees used in this study, as well as the results of IC analyses are available from the figshare repository (https://figshare.com/s/499a7a659dd75d4282cc).

Empirical Data Sets

Three collections of empirical data sets were analyzed in this study. We first compared the performance of quartet-based and bipartitions-based IC scores on bootstrap replicate trees using a data set from Wickett et al. (2014) (referred to as "B200" in our study), including 844 maximum-likelihood gene trees, each associated with 200 bootstrap replicate trees. All trees were downloaded from https://goo.gl/51Sg81. We further randomly sampled 100 and 50 bootstrap replicate trees for each gene and created two additional data sets B100 and B50, respectively. We then applied the IC measures to the 23-taxon yeast data set and the 48-taxon avian data set, which were used in Kobert et al. (2016) to evaluate various adjustment schemes of the original IC/ICA scores. The two data sets were originally published in Salichos and Rokas (2013) and Jarvis et al. (2014), respectively. Gene trees and species trees in these two data sets were downloaded from https://github.com/Kobert/ICTC. We also calculated quartet-based IC scores for the relationships among 42 therian mammals. We retrieved from the OrthoMaM v9 database nucleotide sequence alignments of 5246 coding-exons that are present in at least four species, including one of the three marsupials (*Monodelphis domestica*, *Macropus eugenii*, and *Sarcophilus harrisii*), and the monotreme (*Ornithorynchus anatinus*) was removed as described in Scornavacca and Galtier (2017). Single-gene trees were inferred using IQ-TREE v1.6.5 (Nguyen et al. 2015) under the GTR+G model with 10 searches. A coalescent-based species tree was then estimated from the single-gene trees using ASTRAL v4.11.2 (Mirarab and Warnow 2015).

Calculation of Branch Support and IC Scores

For a given reference tree and evaluation tree set, the branch support values (gene support frequencies or bootstrap support values) for internal branches in the reference tree were calculated using RAxML v8.2.10 with the "-f b" option. Similarly, the bipartition-based IC scores were calculated by using RAxML with the "-f i" option. The original IC/ICA scores were reported if all evaluation trees were complete, whereas the PIC/PICA and LIC/LICA scores (IC/ICA scores adjusted under the "Probabilistic" and "Lossless" schemes, respectively) were reported if some evaluation trees were partial. The underlying bipartition frequencies for calculating the IC/ICA scores were obtained by turning on the "-C" option in RAxML. The quartet-based IC scores were calculated using the program *QuartetScores*, and the LQ-IC/QP-IC/EQP-IC scores were always reported regardless of the status of missing taxa in the evaluation tree set. In cases where the reference tree and evaluation trees are species tree and gene trees, respectively, the LPP supports were calculated using ASTRAL v4.11.2.

Comparing the Performance of Quartet-Based and Bipartition-Based IC Measures

Robustness to missing data.—Here, we define the "robustness" of an IC measure as the distance between the IC scores calculated from evaluation tree sets before and after taxon-/gene-pruning, similar to "accuracy" as defined in Kobert et al. (2016). The quartet-based and bipartition-based IC scores were first calculated for each of the G1000_Original/L1/L2/L3/E1/E2 data sets, as well as the gene-pruned data sets derived from them. Then, for each type of IC scores and each reference tree, pairwise distances were calculated between the G1000_Original data set and each of the taxon-/gene-pruned data sets. The robustness was measured by pairwise distance instead of Spearman (or Pearson) correlation coefficient because two sets of very different scores can still have very high correlation coefficient (e.g., scores in one set are one-tenth of the corresponding scores in the other set). However, unlike in Kobert et al. (2016), here we used the Euclidean distance:

$$D = \sqrt{\sum_{i=1}^{n} (IC_i - IC'_i)^2}$$

where n is the number of internal branches in the reference tree (n = 98 for these simulated data sets), IC_i and IC'_i refer to the IC scores based on the G1000_Original and the pruned data sets, respectively, for the same internal branch i. For easier interpretation, we normalized the Euclidean distances by the largest possible Euclidean distance between two sets of IC scores on the reference tree (e.g., for the simulated data sets where each reference tree contains 98 internal branches, the largest possible distance is $\sqrt{98 * 2^2} \approx 19.80$). In addition, in each pairwise comparison, we also calculated the fraction of internal branches for which the IC scores were overestimated (by more than 0.05) on the pruned data set compared with the G1000_Original data set. Similarly, we also evaluated the robustness of various IC measures on bootstrap replicate trees by comparing the scores based on the B200 data set against those based on the B100 or B50 data sets.

Random evaluation tree test.—In this test, the topologies of all evaluation trees in the G1000_Original/E1/E2 data sets were randomized. Since the randomized evaluation trees have the same sets of taxa as the original trees, the pattern of missing taxa in each data set was kept the same. The quartet-based and bipartition-based IC scores

were then calculated from the original reference trees and randomized evaluation trees.

Altered reference tree test.—Here, the topologies of the reference trees were altered, whereas the evaluation tree topologies remained unchanged. First, we calculated the rRF distance (Robinson 1971) between each evaluation tree in the G1000_Original data set and its corresponding reference tree. Polytomous evaluation trees were randomly resolved before calculating rRF distances. Second, we classified the evaluation trees in to five categories based on their rRF distances; the ranges of rRF distances of the five categories were: [0.1, 0.3), [0.3, 0.5), [0.5, 0.7), [0.7, 0.9), and [0.9, 1]. Finally, we randomly selected 10 evaluation trees from each category to be the new reference trees. The evaluation tree sets to which the new reference trees belonged were also selected as the new evaluation tree sets. Similarly, for data sets G1000_E1/E2, the evaluation tree sets that match with the new reference trees were selected for this test. The quartet-based and bipartition-based IC scores were then calculated from the new, altered reference trees and their corresponding evaluation trees.

Test of Quartet-Based IC Measures in Predicting Branch Correctness

The performance of quartet-based IC measures for predicting branch correctness was tested on the G1000_E2 data set. We first used the 1000 taxon-pruned gene trees in each of the 50 replicates to estimate a species tree using ASTRAL. Internal branches in each of the 50 estimated species trees were classified as true (or false) if the associated bipartition is present (or absent) in the corresponding true species tree. We then calculated the quartet-based IC scores and LPP supports for all internal branches in the estimated species trees. For each measure, an internal branch is predicted to be correct if the associated score (or support) is above a certain threshold. Lastly, for each measure, we calculated (using varying threshold values) the following three statistics which are commonly used to assess the performance of branch support methods (Anisimova et al. 2011; Sayyari and Mirarab 2016):

$$\text{FPR} = \frac{\text{false internal branches prediced to be correct}}{\text{false internal branches}}$$

$$\text{recall} = \frac{\text{true internal branches predicted to be correct}}{\text{true internal branches}}$$

$$\text{precision} = \frac{\text{true internal branches predicted to be correct}}{\text{internal branches predicted to be correct}}$$

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.440874g.

Funding

This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University. X.Z. was supported by the National Key Project for Basic Research of China (973 Program, No. 2015CB150600) and the open fund from Key Laboratory of Ministry of Education for Genetics, Breeding and Multiple Utilization of Crops, College of Crop Science, Fujian Agriculture and Forestry University (GBMUC-2018-005). A.R. was supported by the National Science Foundation (DEB-1442113) and a Guggenheim fellowship. Part of this work was financially supported by the Klaus Tschira Foundation.

Acknowledgments

We thank Olivier Gascuel, Mike Steel, two anonymous reviewers, and members of the Rokas lab for constructive comments that greatly improved the manuscript.

References

Aberer A.J., Stamatakis A. 2011. A simple and accurate method for rogue taxon identification. IEEE International Conference on Bioinformatics and Biomedicine. Atlanta (GA): IEEE. p. 118–122.

Anisimova M., Gil M., Dufayard J.F., Dessimoz C., Gascuel O. 2011. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. Syst. Biol. 60:685–699.

Avni E., Cohen R., Snir S. 2015. Weighted quartets phylogenetics. Syst. Biol. 64:233–242.

Castoe T.A., de Koning A.P., 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.

Chen M.Y., Liang D., Zhang P. 2015. Selecting question-specific genes to reduce incongruence in phylogenomics: a case study of jawed vertebrate backbone phylogeny. Syst. Biol. 64:1104–1120.

Chesters D. 2017. Construction of a species-level tree of life for the insects and utility in taxonomic profiling. Syst. Biol. 66:426–439.

Chifman J., Kubatko L. 2014. Quartet inference from SNP data under the coalescent model. Bioinformatics. 30:3317–3324.

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.

Driskell A.C., Ane C., Burleigh J.G., McMahon M.M., O'Meara B.C., Sanderson M.J. 2004. Prospects for building the tree of life from large sequence databases. Science. 306:1172–1174.

Esselstyn J.A., Oliveros C.H., Swanson M.T., Faircloth B.C. 2017. Investigating difficult nodes in the placental mammal tree with expanded taxon sampling and thousands of ultraconserved elements. Genome Biol. Evol. 9:2308–2321.

Fernandez R., Kallal R.J., Dimitrov D., Ballesteros J.A., Arnedo M.A., Giribet G., Hormiga G. 2018. Phylogenomics, diversification dynamics, and comparative transcriptomics across the spider tree of life. Curr. Biol. 28:2190–2193.

Foley N.M., Springer M.S., Teeling E.C. 2016. Mammal madness: is the mammal tree of life not yet resolved? Philos. Trans. R. Soc. Lond. B Biol. Sci. 371:20150140.

Gadagkar S.R., Rosenberg M.S., Kumar S. 2005. Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. J. Exp. Zool. B Mol. Dev. Evol. 304:64–74.

Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y., 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., Prosdocimi F., Samaniego J.A., Vargas Velazquez A.M., Alfaro-Nunez A., Campos P.F., Petersen B., Sicheritz-PontenT., 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., Zhang G. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science. 346:1320–1331.

Jeffroy O., Brinkmann H., Delsuc F., Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22:225–231.

Kobert K., Salichos L., Rokas A., Stamatakis A. 2016. Computing the internode certainty and related measures from partial gene trees. Mol. Biol. Evol. 33:1606–1617.

Krabberod A.K., Orr R.J.S., Brate J., Kristensen T., Bjorklund K.R., Shalchian-Tabrizi K. 2017. Single cell transcriptomics, mega-phylogeny, and the genetic basis of morphological innovations in Rhizaria. Mol. Biol. Evol. 34:1557–1573.

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.

Leveille-Bourret E., Starr J.R., Ford B.A., Lemmon E.M., Lemmon A.R. 2017. Resolving rapid radiations within angiosperm families using anchored phylogenomics. Syst. Biol. 67:94–112.

Li Z., Defoort J., Tasdighian S., Maere S., Van de Peer Y., De Smet R. 2016. Gene duplicability of core genes is highly consistent across all angiosperms. Plant Cell. 28:326–344.

Maddison W.P. 1997. Gene trees in species trees. Syst. Biol. 46: 523–536.

Mirarab S., Reaz R., Bayzid M.S., Zimmermann T., Swenson M.S., Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics. 30:i541–i548.

Mirarab S., Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics. 31:i44–i52.

Misof B., Liu S., Meusemann K., Peters R.S., Donath A., Mayer C., Frandsen P.B., Ware J., Flouri T., Beutel R.G., Niehuis O., Petersen M., Izquierdo-Carrasco F., Wappler T., Rust J., Aberer A.J., Aspock U., Aspock H., Bartel D., Blanke A., Berger S., Bohm A., Buckley T.R., Calcott B., Chen J., Friedrich F., Fukui M., Fujita M., Greve C., Grobe P., Gu S., Huang Y., Jermiin L.S., Kawahara A.Y., Krogmann L., Kubiak M., Lanfear R., Letsch H., Li Y., Li Z., Li J., Lu H., Machida R., Mashimo Y., Kapli P., McKenna D.D., Meng G., Nakagaki Y., Navarrete-Heredia J.L., Ott M., Ou Y., Pass G., Podsiadlowski L., Pohl H., von Reumont B.M., Schutte K., Sekiya K., Shimizu S., Slipinski A., Stamatakis A., Song W., Su X., Szucsich N.U., Tan M., Tan X., Tang M., Tang J., Timelthaler G., Tomizuka S., Trautwein M., Tong X., Uchifune T., Walzl M.G., Wiegmann B.M., Wilbrandt J., Wipfler B., Wong T.K., Wu Q., Wu G., Xie Y., Yang S., Yang Q., Yeates D.K., Yoshizawa K., Zhang Q., Zhang R., Zhang W., Zhang Y., Zhao J., Zhou C., Zhou L., Ziesmann T., Zou S., Li Y., Xu X., Zhang Y., Yang H., Wang J., Wang J., Kjer K.M., Zhou X. 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science. 346:763–767.

Nagy L.G., Ohm R.A., Kovacs G.M., Floudas D., Riley R., Gacser A., Sipiczki M., Davis J.M., Doty S.L., de Hoog G.S., Lang B.F., Spatafora J.W., Martin F.M., Grigoriev I.V., Hibbett D.S. 2014. Latent homology and convergent regulatory evolution underlies the repeated emergence of yeasts. Nat. Commun. 5:4471.

Nesbo C.L., Boucher Y., Doolittle W.F. 2001. Defining the core of nontransferable prokaryotic genes: the euryarchaeal core. J. Mol. Evol. 53:340–350.

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.

Nieselt-Struwe K., von Haeseler A. 2001. Quartet-mapping, a generalization of the likelihood-mapping procedure. Mol. Biol. Evol. 18:1204–1219.

Pease J.B., Brown J.W., Walker J.F., Hinchliff C.E., Smith S.A. 2018. Quartet sampling distinguishes lack of support from conflicting support in the green plant tree of life. Am. J. Bot. 105:385–403.

Ranwez V., Gascuel O. 2001. Quartet-based phylogenetic inference: improvements and limits. Mol. Biol. Evol. 18:1103–1116.

Robinson D.F. 1971. Comparison of labeled trees with valency three. J. Comb. Theory B 11:105–119.

Salichos L., Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature. 497:327–331.

Salichos L., Stamatakis A., Rokas A. 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. Mol. Biol. Evol. 31:1261–1271.

Sand A., Holt M.K., Johansen J., Brodal G.S., Mailund T., Pedersen C.N. 2014. tqDist: a library for computing the quartet and triplet distances between binary or general trees. Bioinformatics. 30:2079–2080.

Sayyari E., Mirarab S. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. Mol. Biol. Evol. 33:1654–1668.

Scornavacca C., Galtier N. 2017. Incomplete lineage sorting in mammalian phylogenomics. Syst. Biol. 66:112–120.

Shannon C.E. 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27:379–423.

Shen X.X., Opulente D.A., Kominek J., Zhou X., Steenwyk J.L., Buh K.V., Haase M.A.B., Wisecaver J.H., Wang M., Doering D.T., Boudouris J.T., Schneider R.M., Langdon Q.K., Ohkuma M., Endoh R., Takashima M., Manabe R.I., Cadez N., Libkind D., Rosa C.A., DeVirgilio J., Hulfachor A.B., Groenewald M., Kurtzman C.P., Hittinger C.T., Rokas A. 2018. Tempo and mode of genome evolution in the budding yeast subphylum. Cell. 175:1533–1545.e20.

Shen X.X., Zhou X., Kominek J., Kurtzman C.P., Hittinger C.T., Rokas A. 2016. Reconstructing the backbone of the saccharomycotina yeast phylogeny using genome-scale data. G3 (Bethesda). 6:3927–3939.

Slowinski J.B., Page R.D. 1999. How should species phylogenies be inferred from sequence data? Syst. Biol. 48:814–825.

Smith S.A., Moore M.J., Brown J.W., Yang Y. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. BMC Evol. Biol. 15:150.

Song S., Liu L., Edwards S.V., Wu S. 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc. Natl. Acad. Sci. USA. 109:14942–14947.

Steenwyk J.L., Shen X.X., Lind A.L., Goldman G.G., Rokas A. 2019. A robust phylogenomic time tree for biotechnologically and medically important fungi in the genera Aspergillus and Penicillium. mBio. 10:e00925–19.

Strimmer K., von Haeseler A. 1996. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. Mol. Biol. Evol. 13:964–964.

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.

Tarver J.E., Dos Reis M., Mirarab S., Moran R.J., Parker S., O'Reilly J.E., King B.L., O'Connell M.J., Asher R.J., Warnow T., Peterson K.J., Donoghue P.C., Pisani D. 2016. The interrelationships of placental mammals and the limits of phylogenetic inference. Genome Biol. Evol. 8:330–344.

Wang Y., Zhou X., Yang D., Rokas A. 2015. A genome-scale investigation of incongruence in culicidae mosquitoes. Genome Biol. Evol. 7:3463–3471.

Wickett N.J., Mirarab S., Nguyen N., Warnow T., Carpenter E., Matasci N., Ayyampalayam S., Barker M.S., Burleigh J.G., Gitzendanner M.A., Ruhfel B.R., Wafula E., Der J.P., Graham S.W., Mathews S., Melkonian M., Soltis D.E., Soltis P.S., Miles N.W., Rothfels C.J., Pokorny L., Shaw A.J., DeGironimo L., Stevenson D.W., Surek B., Villarreal J.C., Roure B., Philippe H., dePamphilis C.W., Chen T., Deyholos M.K., Baucom R.S., Kutchan T.M., Augustin M.M., Wang J., Zhang Y., Tian Z., Yan Z., Wu X., Sun X., Wong G.K., Leebens-Mack J. 2014. Phylotranscriptomic analysis of the origin

and early diversification of land plants. Proc. Natl. Acad. Sci. USA. 111:E4859–E4868.
Wilkinson M. 2006. Identifying stable reference taxa for phylogenetic nomenclature. Zool. Scr. 35:109–112.
Yang Y., Moore M.J., Brockington S.F., Mikenas J., Olivieri J., Walker J.F., Smith S.A. 2018. Improved transcriptome sampling pinpoints 26 ancient and more recent polyploidy events in Caryophyllales, including two allopolyploidy events. New Phytol. 217:855–870.
Yang Y., Moore M.J., Brockington S.F., Soltis D.E., Wong G.K., Carpenter E.J., Zhang Y., Chen L., Yan Z., Xie Y., Sage R.F., Covshoff S., Hibberd J.M., Nelson M.N., Smith S.A. 2015. Dissecting molecular evolution in the highly diverse plant clade caryophyllales using transcriptome sequencing. Mol. Biol. Evol. 32:2001–2014.
Zhaxybayeva O., Gogarten J.P. 2002. Bootstrap, Bayesian probability and maximum likelihood mapping: exploring new tools for comparative genome analyses. BMC Genomics 3:4.
Zhaxybayeva O., Gogarten J.P., Charlebois R.L., Doolittle W.F., Papke R.T. 2006. Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. Genome Res. 16:1099–1108.
Zhong B., Liu L., Yan Z., Penny D. 2013. Origin of land plants using the multispecies coalescent model. Trends Plant Sci. 18:492–495.