Appendix S1: Supplementary Methods and Results: A holistic approach to phylogenetic branch evaluation on large trees

James B. Pease[1*], Joseph W. Brown[2], Joseph F. Walker[2], Cody E. Hinchliff[3], and Stephen A. Smith[2*]

[1]Department of Biology, Wake Forest University, 455 Vine Street, Winston-Salem, North Carolina 27101, USA
[2]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan 48109, USA
[3]University of Idaho Dept. of Biological Sciences. 875 Perimeter Drive, MS 3051, Moscow, Idaho 83844, USA
[*]Corresponding Authors: peasejb@wfu.edu, eebsmith@umich.edu

Contents

1 Description of the Quartet Sampling test **1**
1.1 Quartet Concordance (QC) 2
1.2 Quartet Differential (QD) 3
1.3 Quartet Informativeness (QI) 3
1.4 Quartet Fidelity (QF) 3

2 Partitioned simulations and empirical tests **4**

3 ILS and Introgression simulations **4**

4 Comparison of RogueNaRok and Quartet Fidelity **5**

5 Table: Comparison of Rogue Taxon RBIC scores and QF scores **6**

6 Table: Quartet Sampling component means and interquartile ranges from empirical datasets tested **7**

Supplementary References **8**

1 Description of the Quartet Sampling test

The Quartet Sampling (QS) procedure was originally outlined by Hinchliff and Smith (2014), but is presented here with modifications. In this procedure, for a given bifurcating tree (T), each

observed internal edge (b) partitions the set of all tips (S) in T into four non-overlapping subsets $S_{j\in\{1,2,3,4\}}$ (Fig. 1). For each of a specified number of sampling replicates (r), a single taxon is sampled randomly from each S_j. Note that there are $\prod_{i=1}^{4} S_i$ total possible unique quartets for each internal branch. Aligned sequences from these four quartets are extracted, and the likelihoods for each of the three possible unrooted topological resolutions for the four sampled taxa are calculated. In this study, likelihoods were calculated with RAxML (Stamatakis, 2014) and PAUP* (Swofford and Sullivan, 2003), but the procedure is adaptable to any likelihood calculator.

We refer to the quartet topology that includes b as t_1 (where t_1 is topologically concordant with T) and label the two alternatives (discordant with T) arbitrarily as t_2 and t_3.The associated likelihood values are labeled L_1, L_2, and L_3, respectively. The topology with the highest likelihood value and the difference between the best and second-best likelihood values (ΔL) are recorded. If ΔL exceeds a specified minimum likelihood cutoff (k), then the replicate is tallied among the counts for each topology on the given branch (n_1, n_2, and n_3, respectively). Alternatively, if $\Delta L < k$, then the replicate is tallied under n_u ("uncertain") where $r = n_1 + n_2 + n_3 + n_u$.

1.1 Quartet Concordance (QC)

Across all replicates, we then calculate a "quartet concordance" score (QC) indicating the degree to which n_1 occurs more often than the two alternative resolutions for focal branch b (Eq. 1).

$$\text{QC}_b = \log_z(z) + \sum_{i\in\{1,2,3\}} \frac{n_i}{N} \log_z(\frac{n_i}{N}) \tag{1}$$

Here, $N = r - n_u$ (i.e., the total number of replicates where the optimal resolution exceeded the minimum likelihood threshold ΔL) and z is the total number of non-zero counts among the three candidate topologies. In the case of quartets, $z = 2$ when two alternative topologies are counted and $z = 3$ when all three alternatives are sampled. The QC score is an application of Shannon's entropy (Shannon, 1948), similar to the ICA score (Salichos et al., 2014), but applied only for the given set of quartets on a branch.Note that, like (Salichos et al., 2014), we reverse the sign of conventional Shannon entropy and add 1 ($\log_z(z)$) to reflect that QC is a measure of support rather than a lack thereof. QC takes on the range $[-1, 1]$ gives a measure of how much more frequently concordant t_1 occurs compared to the discordant t_2 and t_3. When QC ≈ 0 there is no confidence in t_1; this occurs when all non-zero topology counts are equal. QC approaches 1 when $n_1 > 0$ and $n_2 \approx n_3 \to 0$, meaning the vast majority of sampled quartets support b on T. Conversely, QC < 0 means an alternative topology is more strongly supported, with QC $= -1$ being the most extreme case when a single alternative topology is sampled. Strictly speaking QC is undefined when only one of the three topologies has a non-zero count. In these instances QC is set to 1 (if $n_1 > 0$ and $n_2 = n_3 = 0$) to reflect perfect support or -1 (if $n_2 > 0$ and $n_1 = n_3 = 0$) to reflect perfect conflict.

Additionally, we can define a tree confidence QTC score (analogous the TC score from Salichos et al. 2014; Eq. 2):

$$\text{QTC} = \sum_{b\in S} \text{QC}_b \tag{2}$$

This score could be used to compare total QC values for different trees using the same alignment data, though this is not implemented in the present study.

1.2 Quartet Differential (QD)

The QS procedure also calculates the discordant "Quartet Differential" (QD), which measures the difference in count of the two discordant topologies for focal branch b (Eq. 3):

$$\text{QD}_b = 1 - \frac{|n_2 - n_3|}{n_2 + n_3} \tag{3}$$

and takes on the range $[0, 1]$. This measure is analogous in function and interpretation to the numerator of the f- and D-statistics for introgression (Durand et al., 2011; Green et al., 2010; Reich et al., 2009; Pease and Hahn, 2015), but using counts of tree topologies rather than individual site counts or site frequencies. All of these measures are based on the model that, under a stochastic process of incomplete lineage sorting alone, the proportion of the two discordant topologies should be approximately equivalent, such that $n_2 \approx n_3$ and therefore QD $\approx$ 1. When introgression is present, one discordant topology that is the result of an introgression event will be enriched in frequency relative to the other (i.e., $0 \leq$ QD < 1). Note that this scale is inverted from the usual D-statistic measure in order for the score of 1 to indicate full support (as is typical in phylogenetics).

As a simple relative ratio, we note that the QD score (as with D-statistic) can take on extreme values when the counts of discordant trees (n_2 and n_3) are low. Therefore, the QD score is best evaluated in concert with the QC score. A QD score of 0.1 is far less meaningful when QC = 0.95 (few discordant trees sampled) versus QC = 0.3 (many discordant trees sampled).

1.3 Quartet Informativeness (QI)

The "Quartet Informativeness" (QI) score is calculated only when a likelihood differential is used. For a given likelihood cutoff k, a quartet is considered "informative" if the likelihood of the best likelihood tree (of the three possible trees) differs in ln-likelihood from the second-best likelihood tree by a value of $\Delta L \leq k$. "Uninformative" Trees that fail the likelihood cutoff are placed in a separate count (n_u). QI is then calculated for focal branch b as (Eq. 4):

$$\text{QI}_b = \frac{n_1 + n_2 + n_3}{n_1 + n_2 + n_3 + n_u} \tag{4}$$

and takes on the range $[0, 1]$.

1.4 Quartet Fidelity (QF)

In addition to describing relative support on each internal branch, the QS method also provides a means to report unreliable terminal branches (a.k.a. "rogue taxa"; Wilkinson 1996). The "quartet fidelity" score (QF) measures a consistently significant decrease in the reliability of tree inference

when a given "rogue taxon" is included. For each tip (x) in the tree, the number of times that x supports the concordant tree ($n_{(x)1}$) or each of the discordant trees ($n_{(x)2}$ and $n_{(x)3}$; Eq. 5):

$$\text{QF}_x = \frac{n_{(x)1}}{n_{(x)1} + n_{(x)2} + n_{(x)3}} \tag{5}$$

The QF score is a proportion on a $[0, 1]$ scale. Unsuccessful replicates (failing likelihood cutoff) are not included.

2 Partitioned simulations and empirical tests

We used the same procedure as in the main methods to simulated trees with 500 tips and associated alignment with ten nucleotide partitions, each with 500 sites under the Jukes-Cantor model. Both single partition and multiple partition datasets were generated for the 100 tip simulations. To reflect this, each partition was assigned a rate of evolution from slow (maximum branch length = 0.1) to fast (maximum branch length = 2), and the underlying tree was scaled accordingly before it was supplied to INDELible v. 1.03 (Fletcher and Yang, 2009). In addition to the variation in partition-specific rates, we also produced a sparse alignment by subsampling the partition-specific tips phylogenetically. We then did a weighted random subsampling with more sequence retained in fast partitions than from slow ones, and then preorder-traversed the true tree. The number of taxa to retain within each clade was proportionately determined based on the size clade. About 0.5% of the taxa in slow partitions were retained, as were about 99.5% of taxa in fast partitions. On average, the resulting multi-partition datasets had 50% missing data. Overall this multi-partition data was qualitatively consistent with the results presented in the main methods and partitioning in the main datasets also did not substantially change the results of the empirical testing.

3 ILS and Introgression simulations

A simulated tree (Appendix S5) was generated using the `pxbdsim` function in `phyx` (Brown et al., 2017). Note that the procedure below was tested using several starting trees with qualitatively similar results.

```
pxbdsim -e 20 > simtree.tre
```

Then this tree was used as the input for a Python script that first scales by multiplying all branch lengths in the tree using the scaling parameter (`-s`). Then alignments are generated using *ms* (Hudson, 2002) simulates 20 loci (`-l 20`), of 5000 bp (`-w 5000`) using the population parameters of $N_e = 10^5$ (`-N 100000`) and mutation rate per site $\mu = 10^{-8}$ per site per generation (`-u 1e-8`). The alignment was converted from *ms* substitution data by a simple Kimura two-parameter implementation where nucleotides and the substituted nucleotides are drawn at random conditioned on an initial GC content and a transition:transversion ratio of 2:1. RAxML (Stamatakis, 2014) was then run on the resulting alignment to infer a tree (arbitrarily rooted with "taxon_20") (`-g taxon_20`). Next QS was run for 200 replicates (`-q 200`) on the ML-tree (not the original tree) and the generated alignment. This procedure was run for 100

replicates (`-n 100`), where QC and QD values were tabulated for each replicate for each branch that was congruent with the original tree.

```
python3 qstreesim.py simtree.tre -o $OUTPUT -u 1e-8 -N 100000 \
-w 5000 \ -l 20 -n 100 -s $SCALING \
--qs_exec $QSPATH/quartet_sampling.py -q 200 -g taxon_20
```

This procedure was repeated using the scaling parameters: 10, 7, 5, 4, 3, 2.5, 2, 1.5, 1, and 0.5. The shortening of branch lengths increases the amount of ILS.

Additionally, we ran the same simulation with the same input tree this time fixing the scaling parameter at `-s 2` and 0.5 was added to each branch (`-a 0.5`) to minimize background ILS. All other parameters were the same except that migration was turned on between "taxon_7" and "taxon_6" (`-S 7 -D 6`) from 0 to 0.5 coalescent time units (`-T 0.0 0.5`) at varying rates of migration (`-M`). QS was performed this using on the original tree (`--orig-tree`) and the simulated alignments.

```
python3 qstreesim.py simtree.tre -o $OUTPUT -u 1e-8 -N 100000 \
-w 5000 -l 20 -n 100 -a 0.5 -s 2 \
--qs_exec $QSPATH/quartet_sampling.py -q 200 -g taxon_20 \
-S 7 -D 6 -T 0.0 0.5 -M $MIGRATION --orig-tree
```

This procedure was repeated using the migration rate parameters: 0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4. The increasing migration rate simulated increasing introgression.

4 Comparison of RogueNaRok and Quartet Fidelity

We compared the QF scores to the results of RogueNaRok (Aberer et al., 2012) for identification of "rogue taxa." RogueNaRok works by using bootstrap replicates to provide a relative bipartition information criterion (RBIC), and identifying taxa that can be removed to improve the RBIC. We compared this to the QF, which represents how often a taxon in a quartet is concordant with the species tree topology. Using data from Zanne et al. (2014) as an example of a large sparse data matrix, we found an overall RBIC score of 59.05% and 3433 taxa were suggested as being able to improve the RBIC score if removed. The Zanne et al. (2014) dataset contained a mean score of 0.15 indicating that many of taxa had alternate topologies. The RogueNaRok aided in identifying taxa that could be removed to help stabilize the results, for which QF does not have a defined cutoff.

When examining the Walker et al. (2017) phylogenomic dataset, we found the RogueNaRok analysis provided a perfect RBIC score and the mean QF score was 0.89. The QF score, showed low support for *Drosophyllum lusitanicum*, a taxon previously shown to move around in phylogenomic analyses (Walker et al., 2017). This indicates that in the case of phylogenomic datasets, where perfect bootstrap support is often found, the QF score can be used to help identify rogue taxa.

In (Pease et al., 2016) (Supplementary Table 1.26) RogueNaRok was used to compute the taxonomic instability index (TII) using gene trees and 100kb partition trees with similar results to QF.

5 Table: Comparison of Rogue Taxon RBIC scores and QF scores

Dataset	Initial RBIC (# of taxa suggested to be removed)	QF Mean
Cannon et al. (2015)	0.959	0.69
Pryer et al. (2016)	0.877	0.15
Walker et al. (2017)	1.00000	0.89
Xi et al. (2014)	0.982558	0.74
Zanne et al. (2014)	0.590588 (3433)	0.15

6 Table: Quartet Sampling component means and interquartile ranges from empirical datasets tested

Dataset	# Taxa	QC mean (IQR)	QD mean (IQR)	QI mean (IQR)	QF mean (IQR)
Hinchliff and Smith (2014)	13903	0.16 (–0.15,0.49)	0.39 (0.05,0.67)	0.35 (0.2,0.47)	0.21 (0.13,0.26)
Zanne et al. (2014)	31749	0.2 (–0.11,0.63)	0.35 (0,0.67)	0.15 (0.06,0.22)	0.15 (0.07,0.2)
Wickett et al. (2014)	103	0.55 (0.21,1.0)	0.22 (0.0,0.29)	0.98 (0.97,1.0)	0.76 (0.69,84)
Xi et al. (2014)	46	0.55 (0.18,1)	0.47 (0.18,0.83)	0.98 (0.96,1.0)	0.74 (0.68,0.78)
Pryer et al. (2016)	34	0.71 (0.49,1.0)	0.25 (0.0,0.47)	0.81 (0.66,1.0)	0.74 (0.69,0.8)
Cannon et al. (2015)	37	0.51 (0.18,0.92)	0.29 (0.0,0.52)	0.93 (0.91,1.0)	0.69 (0.65,0.75)
Pease et al. (2016)	30	0.15 (–0.13,0.46)	0.33 (0.0,0.67)	0.35 (0.19,0.47)	0.20 (0.13,0.27)
Walker et al. (2017)	13	0.85 (1,1)	0.16 (0.08,0.24)	1.0 (1.0,1.0)	0.89 (0.86,0.89)

Table 1: Quartet Sampling component means and interquartile ranges from empirical datasets tested.

References

Aberer, A. J., D. Krompass, and A. Stamatakis. 2012. Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. *Systematic Biology* 62: 162–166.

Brown, J. W., J. F. Walker, and S. A. Smith. 2017. Phyx: phylogenetic tools for unix. *Bioinformatics* 33: 1886–1888.

Cannon, S. B., M. R. McKain, A. Harkess, M. N. Nelson, S. Dash, M. K. Deyholos, Y. Peng, et al. 2015. Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. *Molecular Biology and Evolution* 32: 193–210.

Durand, E. Y., N. Patterson, D. Reich, and M. Slatkin. 2011. Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution* 28: 2239–2252.

Fletcher, W. and Z. Yang. 2009. INDELible: a flexible simulator of biological sequence evolution. *Molecular Biology and Evolution* 26: 1879–1888.

Green, R. E., J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, et al. 2010. A draft sequence of the neandertal genome. *Science* 328: 710–722.

Hinchliff, C. E. and S. A. Smith. 2014. Some limitations of public sequence data for phylogenetic inference (in plants). *PLoS ONE* 9: e98986.

Hudson, R. R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18: 337–338.

Pease, J. B., D. C. Haak, M. W. Hahn, and L. C. Moyle. 2016. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLoS Biology* 14: e1002379.

Pease, J. B. and M. W. Hahn. 2015. Detection and polarization of introgression in a five-taxon phylogeny. *Systematic Biology* 64: 651–662.

Pryer, K. M., L. Huiet, F.-W. Li, C. J. Rothfels, and E. Schuettpelz. 2016. Maidenhair ferns (*Adiantum*) are indeed monophyletic and sister to shoestring ferns, vittarioids (Pteridaceae). *Systematic Botany* 41: 17–23.

Reich, D., K. Thangaraj, N. Patterson, A. L. Price, and L. Singh. 2009. Reconstructing Indian population history. *Nature* 461: 489–494.

Salichos, L., A. Stamatakis, and A. Rokas. 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Molecular Biology and Evolution* 31: 1261–1271.

Shannon, C. E. 1948. A note on the concept of entropy. *Bell System Technical Journal* 27: 379–423.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313.

Swofford, D. L. and J. Sullivan. 2003. Phylogeny inference based on parsimony and other methods using PAUP*. *In* The Phylogenetic Handbook: a Practical Approach to Phylogenetic Analysis and Hypothesis Testing, volume 7. Cambridge University Press. ISBN 0521730716, 160–206.

Walker, J. F., Y. Yang, M. J. Moore, J. Mikenas, A. Timoneda, S. F. Brockington, and S. A. Smith. 2017. Widespread paleopolyploidy, gene tree conflict, and recalcitrant relationships among the carnivorous Caryophyllales. *American Journal of Botany* 104: 858–867.

Wickett, N. J., S. Mirarab, N. Nguyen, T. Warnow, E. Carpenter, N. Matasci, S. Ayyampalayam, et al. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences of the United States of America* 111: E4859–E4868.

Wilkinson, M. 1996. Majority-rule reduced consensus trees and their use in bootstrapping. *Molecular Biology and Evolution* 13: 437–444.

Xi, Z., L. Liu, J. S. Rest, and C. C. Davis. 2014. Coalescent versus concatenation methods and the placement of *Amborella* as sister to water lilies. *Systematic Biology* 63: 919–932.

Zanne, A. E., D. C. Tank, W. K. Cornwell, J. M. Eastman, S. A. Smith, R. G. FitzJohn, D. J. McGlinn, et al. 2014. Three keys to the radiation of angiosperms into freezing environments. *Nature* 506: 89–92.