Quartet Sampling distinguishes lack of support from conflicting support in the plant tree of life


Department of Biology, Wake Forest University, 455 Vine Street, Winston-Salem, North Carolina 27101, USA

Department of Ecology and Evolutionary Biology, University of Michigan, 830 North University, Ann Arbor, Michigan 48109, USA

Department of Biological Sciences, University of Idaho, 875 Perimeter Drive, MS 3051, Moscow, Idaho 83844, USA

Corresponding Authors: peasejb@wfu.edu, eebsmith@umich.edu

Short Title: Quartet Sampling of discordance in the plant tree of life

1Manuscript received ______; revision accepted ______.
ABSTRACT

Premise of the Study— Phylogenetic support has been difficult to evaluate within the plant tree of life partly due to the inability of standard methods to distinguish conflicted versus poorly informed branches. As phylogenomic and broad-scale datasets continue to grow, support measures are needed that are more efficient and informative.

Methods— We describe the Quartet Sampling (QS) method that synthesizes several phylogenetic and genomic analytical approaches into a quartet-based evaluation system. QS rapidly characterizes discordance in large-sparse and genome-wide datasets, overcoming issues of sparse alignment and distinguishing strong conflict from weak support. We test this method with simulations and recent plant phylogenies inferred from variously sized datasets.

Key Results— QS scores decrease in variance with increasing replicates and are not strongly affected by branch depth. Patterns of QS support from different phylogenies leads to a coherent understanding of ancestral branches defining key disagreements, including Ginkgo+cycad, magnoliids+eudicots (excluding monocots), and mosses+liverworts. The relationships of ANA grade angiosperms, major monocot groups, and bryophytes and fern families are found to likely be the result of discordant evolutionary histories, rather than poor information. Also, analyses of phylogenomic data show QS can detect discordance due to introgression.

Conclusions— The QS method represents an efficient and effective synthesis of phylogenetic tests that offer more comprehensive and specific information on branch support than conventional measures. The QS method corroborates growing evidence that phylogenomic investigations that
incorporate discordance testing are warranted to reconstruct the complex evolutionary histories surrounding in particular ANA grade angiosperms, monocots, and non-vascular plants.

Key words: bootstrap; branch support; discordance; introgression; lineage sorting; phylogenetics; phylogenetic methods; phylogenomics; plant tree of life; quartet sampling

INTRODUCTION

Discordance and uncertainty have emerged as a consistent feature throughout the history of our evolving model of the plant tree of life (Crane, 1985; Chase et al., 1993; Palmer et al., 2004; Soltis et al., 2011; Wickett et al., 2014). Furthermore, key points of uncertainty and contention occur at pivotal transitions in the evolution of plant life on earth, such as the development of vascular tissue (Pryer et al., 2001; Steemans et al., 2009; Banks et al., 2011), the rise of seed-bearing plants (Chase et al., 1993; Chaw et al., 1997; Bowe et al., 2000; Qiu et al., 2006; Jiao et al., 2011), and the explosive radiation of flowering plants (Crane, 1985; The Amborella Genome Project, 2013; Goremykin et al., 2015; Taylor et al., 2015; Simmons, 2016; Edwards et al., 2016). Modern phylogenomic datasets, rather than quelling these disagreements, have now repeatedly shown that these phylogenetic conflicts are not only caused by methodological error or data bias, but also are often the result of genuine biological processes. These forces most commonly include incomplete lineage sorting (ILS), introgressive hybridization, and paralog duplication-loss (e.g., Zhong et al., 2013b; Wickett et al., 2014; Zwickl et al., 2014; Yang et al., 2015; Eaton et al., 2016; Pease et al., 2016b; Goulet et al., 2017; Walker et al., In Press). While several methods have been proposed to deal with these issues in the context of species tree inference (e.g., Zwickl and Hillis, 2002; Ogden and Rosenberg, 2006; Shavit Grievink et al., 2015).
(specifically branch support) that distinguishes poorly informed branches from the increasingly common case of multiple strongly supported, but mutually exclusive, phylogenetic histories. Therefore, the development of systems of phylogenetic evaluation designed to utilize and describe discordance (rather than minimize or control for it as error) offers strong promise to provide a more holistic picture of the current state of the plant tree of life.

Quantification of branch support for molecular phylogenies has been proposed under many methods over the last few decades (Felsenstein, 1985; Farris et al., 1996; Larget and Simon, 1999; Anisimova and Gascuel, 2006; Anisimova et al., 2011; Ronquist et al., 2012; Larget, 2013). Historically, one of the most commonly used tests has been the non-parametric bootstrap (NBS; Felsenstein, 1985), which, along with recent variants like the rapid bootstrap (RBS; Stamatakis et al., 2008), resamples the original data with replacement assuming that aligned sites are independent and identically distributed (i.i.d.) samples that approximate the true underlying distribution (Efron, 1992; Felsenstein, 1985). In practice, the assumptions of NBS (in particular site independence) may rarely be met and can deteriorate under a variety of conditions (Felsenstein, 1985; Felsenstein and Kishino, 1993; Hillis and Bull, 1993; Sanderson, 1995; Andrews, 2000; Alfaro, 2003; Cummings et al., 2003). More recently the UltraFast bootstrap approximation (UFboot) method, utilizing a likelihood-based candidate tree testing, was proposed to address speed and score interpretation issues for NBS (Minh et al. (2013); and see comparison in Simmons and Norton (2014)).

The other most common branch support metric has been the Bayesian posterior probability (PP). PP scores are typically calculated from posterior distributions of trees generated using a
Markov chain Monte Carlo (MCMC) sampler and then summarized using a majority-rule consensus tree (e.g., Larget and Simon, 1999; Drummond and Rambaut, 2007; Ronquist et al., 2012; Larget, 2013). The interpretation of PP values is arguably more straightforward than bootstrap proportions, as PP values represent the simple probability that a clade exists in the underlying tree (conditioned on the model of evolution employed, and assuming the data have evolved in a treelike fashion), and do not involve data resampling. The individual and relative performance of PP has been well-documented and generally favorable (Wilcox et al., 2002; Alfaro, 2003; Cummings et al., 2003; Huelsenbeck and Rannala, 2004). However, it has been recognized that in some scenarios (e.g., oversimplified substitution models) PP may give liberal support (Suzuki et al., 2002; Douady et al., 2003; Nylander et al., 2004), and there are indications that PP also may fail under a multi-species coalescent framework with conflicting phylogenies (e.g., Reid et al., 2013). Some studies have noted the disproportionate effects of a few genes in tipping the scales in the case of low information in large datasets (Brown and Thomson, 2016; Shen et al., 2017).

Modern phylogenetic datasets being assembled to model the plant tree of life are currently trending in two directions. Ongoing efforts to expand genetic sampling to as many plant species as possible have produced increasingly species-rich, but data-sparse, alignments (i.e., large-sparse matrices). Meanwhile, the accelerating accretion of new genomes and transcriptomes will continue to deepen genome-wide datasets with millions of aligned sites. Both types of data present particular challenges to both the tractability and interpretation of phylogenetic branch support methods. NBS scores are known to perform poorly for large-sparse matrices (Wiens and Morrill, 2011; Smith et al., 2011; Roure et al., 2012; Hinchliff and Roalson, 2013; Hinchliff and Smith, 2014b), where the sampling procedure generates uninformative pseudo-replicates that
mostly omit informative sites (or consist of mostly missing data). The problem of uninformative pseudo-replicates also applies to alignment jackknifing used on sparse alignments, where alignments are sampled without replacement. Therefore, alignment jackknifing also does not offer a functional solution for large alignments with missing data.

Quantifying phylogenetic support from complete datasets brings unique challenges. By “complete” dataset, we refer to any molecular data that is not a smaller subset of some larger whole, which means not only whole genomes, but also whole transcriptomes, and even whole plastid/mitochondrial genomes. PPs provide an appropriate testing framework and straightforward interpretation, but available Bayesian methods of analysis and computational speeds do not scale to typical phylogenomic datasets. Resampling methods (including NBS) gauge how well the resampled dataset replicate the initial result, but complete datasets are not samples of any larger whole. PP and NBS scores therefore both appear unsuitable for use on large datasets, the former due to feasibility and the latter due to its assumption (Smith et al., 2009; Hinchliff and Smith, 2014b).

A relatively well-studied, but less frequently applied, family of tests are the likelihood ratio tests, such as the approximate likelihood ratio test (aLRT), the Shimodaira–Hasegawa [SH]-like aLRT (SH-aLRT), and the Bayesian-like LRT (bLRT) (Anisimova and Gascuel, 2006; Guindon et al., 2010; Anisimova et al., 2011). These tests conduct an approximated LRT of the alternative nearest-neighbor interchange (NNI) moves at each branch of an optimized tree topology.

SH-aLRT has been shown to perform well (Anisimova et al., 2011; Simmons and Norton, 2014; Simmons and Randle, 2014) and is computationally efficient enough to be run on large datasets in large part because it does not require generation of multiple topology replicates (as in NBS).

Despite multiple implementations in PhyML and RAxML, and being the default support measure...
in FastTree, this test has rarely been used in published studies (Guindon et al., 2010; Price et al., 2010; Stamatakis, 2014).

As phylogenomics has developed over the last decade, a variety of methods have been introduced to factor the increased data and inherent gene tree-species tree conflict. These methods measure the concordance of gene trees (broadly referring to a phylogeny from any sub-sampled genomic region), including the internode and tree certainty scores (IC/TC; Rokas et al., 2003; Salichos et al., 2014; Kobert et al., 2016), Bayesian concordance factors (Ané et al., 2006), and other concordance measures (Allman et al., 2017). These various certainty scores are developed around the central concept of a branch support statistic that measures concordance of various trees with a particular tree hypothesis. This perspective on phylogenetic evaluation offers much in terms of partitioning phylogenetic discordance and analyzing larger alignments more rapidly in a phylogenomic coalescent-based framework, but these methods that address discordance using gene trees may not be as suitable for large-sparse alignments.

Finally, quartet methods—in particular quartet puzzling methods—have been examined extensively in phylogenetics, especially in relation to phylogenetic reconstruction (Strimmer et al., 1997; Strimmer and von Haeseler, 1997; Ranwez and Gascuel, 2001; Chifman and Kubatko, 2014; Mirarab et al., 2014). A quartet support method called “reliability values,” which correlates with NBS values, is calculated as a side effect of the quartet puzzling procedure (Strimmer et al., 1997; Strimmer and von Haeseler, 1997). More recently, quartet procedures have been explored to facilitate sampling of large-sparse alignments (Misof et al., 2013) and as part of coalescent-based quartet inference methods (Gaither and Kubatko, 2016; Sayyari and Mirarab, 2016). These quartet methods benefit not only from the speed advantages of a smaller alignment, but also from the statistical consistency of quartet trees, which avoid complex lineage sorting.
issues that occur with more speciose phylogenies (Rosenberg, 2002; Degnan and Salter, 2005).

Few measures of support (except concordance methods) explicitly operate with the expectation of multiple histories, or fail to distinguish different causes of poor support for a branch in the phylogeny. In practice, this means when a phylogenetic relationship shows poor support under NBS or PP, the specific cause is not apparent since these tests do not distinguish between the case of multiple supported-but-conflicting phylogenetic relationships and the case of simply low information. Being able to distinguish among these causes of low phylogenetic support and to identify branches that have a strong consensus and a strong secondary evolutionary history would provide valuable insight into the plant tree of life (among others; and see also Brown and Lemmon, 2007).

Here, we describe the Quartet Sampling (QS) method (Fig. 1 and Table 1), which blends aspects from many of the evaluation methods described above and leverages the efficiency of quartet-based evaluation. The goal of the QS method is to evaluate phylogenies (particularly large-sparse and genome-wide) by dissecting phylogenetic discordance to distinguish among (1) differences in phylogenetic branch support due to lack of information (the general goal of bootstrap or Bayesian posteriors), (2) poor support due to discordance as a result of lineage sorting or introgression (as in concordance measures), and (3) low support due to particular taxa with poor or conflicted information (i.e., “rogue taxa”; Wilkinson, 1996; Aberer et al., 2012). These various causes of discordance are frequently surveyed separately in many modern phylogenetic and particularly phylogenomic studies (e.g., Xi et al., 2014a; Wickett et al., 2014; Yang et al., 2015; Pease et al., 2016b; Walker et al., In Press), but the QS method provides a unified method for their execution, interpretation, and reporting. Additionally, the QS method offers a viable means to describe branch support in large phylogenies built from sparse
alignments (10,000–30,000 tips with >80% missing data), datasets for which Bayesian analyses are simply not tractable. We also describe how QS enhances analysis of both genome-wide datasets and smaller-scale multi-gene data sets conventionally used in systematics.

In this study, we will (1) describe the features, parameters, and interpretation of the QS method, (2) validate the QS method with simulations that demonstrate its effectiveness, and (3) evaluate and discuss the plant tree of life by testing the QS method on recently published large-sparse and phylogenomic datasets at timescales spanning from Viridiplantae to sub-generic clades. The goal of the re-analysis of these phylogenies is not to compare the methods of tree inference or dataset composition, but instead to use the QS method to collectively analyze these data as a diverse set of alternative phylogenetic hypotheses of the plant trees of life. We show consistently through these datasets that the QS method acts as a more conservative test of branch support that provides greater discrimination between highly supported and poorly supported branches than other support measures. We also show that the QS method can identify branches in a phylogeny where biological conflict is likely occurring, which can inform targeting of more detailed investigations within a clade. Finally, we show several cases where low support for the given phylogenetic relationship and strong counter-support for an alternative relationship together indicate a coherent consensus. In other cases, QS specifically indicate the likely presence of alternative phylogenetic histories, as distinct from low branch support. We hope this study encourages additional discussion, testing, and innovation of new phylogenetic evaluation methods. We also hope it contributes to the broader discussion about moving the plant tree of life beyond the increasingly difficult goal of resolving a single unified ‘Species Tree’ (Hahn and Nakhleh, 2015; Smith et al., 2015), and into a future where the complex “multiverse” of phylogenetic relationships that is manifest throughout the plant tree of life are more fully
explored and appreciated.

MATERIALS AND METHODS

Quartet Sampling— The Quartet Sampling (QS) procedure outlined here was inspired by aspects from several quartet-based and concordance methods, most particularly the process originally outlined by Hinchliff and Smith (2014b). The QS method takes an existing phylogenetic topology (which can be inferred by any method) and a molecular dataset (not necessarily the one that generated the phylogeny) and separately evaluates one or more internal branches on the given phylogeny. The QS method (Fig. 1) was designed to rapidly and simultaneously assess the confidence, consistency, and informativeness of internal tree relationships, concurrent with an analysis of the reliability of each terminal branch.

For a given phylogeny, each observed internal tree branch partitions the tree into four non-overlapping subsets of taxa (Fig. 1A). These four sets of taxa can exist in three possible relationships: the concordant relationship that matches the configuration in the given topology and two alternative discordant configurations. The QS method repeatedly and randomly samples one taxon from each of the four subsets and then evaluates the likelihood all three possible phylogenies given the sequence data for the randomly selected quartet spanning that particular branch.

For each quartet sampled for the focal branch, the likelihood is evaluated (using RAxML or PAUP*; Stamatakis, 2014; Swofford and Sullivan, 2003) for all three possible topologies that these four sampled taxa can take. The quartet topology with the best likelihood is then recorded and tabulated across all replicates, generating a set of counts (across all replicates per branch) for
the concordant and each of the two discordant relationships. If a minimum alignment overlap is specified, then quartets must contain the minimum number of overlapping non-empty sites for all four taxa to be considered suitable for any calculations. Additionally, a parameter of a minimum likelihood differential may be set. If the most-likely topology (of the three) does not exceed the likelihood of the second-most-likely phylogeny by the set threshold then the quartet is considered “uninformative” and tabulated separately. Therefore, for a given internal branch, the QS method generates counts of the three possible topologies (and uninformative replicates) sampled from different quartets of taxa spanning the particular branch.

The QS method uses these resampled quartet tree counts to calculate three scores for each internal branch of the focal tree (Fig. 1B, Table 1, and Appendix S1; see Supplemental Data with this article). The QC (Quartet Concordance) score is an entropy-like measure (similar to the ICA score; Salichos et al. 2014) that quantifies the relative support among the three possible resolutions of four taxa. When the most commonly sampled topology is concordant with the input tree, then the QC takes positive values in the range (0,1]. Thus, QC=1 when all quartet trees are concordant with the focal branch. When one of the discordant topologies is the most commonly resampled quartet, QC takes negative values in the range [–1,0), approaching –1 when all quartet trees are one of the two discordant phylogenies. When support is evenly split among the three alternative topologies (or two if only two of the three possible are sampled), QC equals 0.

The QD (Quartet Differential) score uses the logic of the f- and D-statistics for introgression (Reich et al., 2009; Green et al., 2010; Durand et al., 2011; Pease and Hahn, 2015) and measures the disparity between the sampled proportions of the two discordant topologies (though with gene tree proportions, rather than site frequencies). The QD score does not specifically quantify introgression nor identify introgressing taxa, but does indicate that one alternative relationship is
sampled more often than the other. High values of QD report that there is one clearly preferred
topology among the two discordant topologies, a potential indication of a biased biological
process beyond background lineage sorting on the given branch, including introgression, strong
rate heterogeneity, heterogeneous base compositions, etc. QD varies in the range \([0,1]\) with a
value of 0, meaning no skew in the proportions of discordant trees, and the extreme value of 1,
which means all discordant trees sampled are only from one of the two possible alternative
relationships.

The QU score (Quartet Uncertainty) quantifies the overall proportion of quartets for a given
branch where the tree with the best likelihood value has a likelihood that is not higher than the
next most likely tree by a given differential cutoff. This ensures that replicates are not counted as
being concordant or discordant when the molecular data itself is effectively equivocal on the
topology by all three options having nearly indistinguishable likelihood scores. QU is measured
in the range \([0,1]\), which indicates the proportion of sampled quartets that did not exceed of the
cutoff. A QU of 0 means no “uncertain” quartets (i.e., highly informative data), while a value of 1
indicates 100% of quartets were uncertain (i.e., no information for the given branch). QU reports
the informativeness of the branch, which in conjunction with QD and QC distinguishes between
branches that have low information versus those with conflicting information (i.e., high
discordance).

Finally, for each terminal taxon, a QF (Quartet Fidelity) score is calculated to report what
proportion of the time inclusion of a given taxon in a quartet (across all branches) results in a
discordant quartet topology. QF is therefore similar in approach to a “rogue taxon” test
(Wilkinson, 1996; Aberer et al., 2012). For a given taxon, the QF score is measured in the range
\([0,1]\) as the proportion of quartet topologies involving the taxon that are concordant with the focal
tree branch. Therefore, QF=1 indicates a given taxon always produces concordant topologies
when used in a quartet replicate. QF values approaching zero indicate mostly discordant
topologies involving this taxon, and may indicate poor sequence quality or identity, a
lineage-specific process that is distorting the phylogeny, or that the taxon is significantly
misplaced in the given tree. Note that QF differs specifically from QC/QD/QU by being a
taxon-specific test across internal branch tests rather than a branch-specific test.

Collectively, these four tests represent a means to distinguish the consistency of a branch
(QC), the presence of a secondary evolutionary history (QD), the amount of information
regarding a branch (QU), and the reliability of individual taxa in the tree (QF; Fig. 1B and see
Table 1). Taken together these tests provide a means to disentangle these effects rather than have
them conflated under a summary score as in standard measures of phylogenetic support. For a full
technical description of the QS method, see Appendix S1.

Guidelines for implementation of QS— We implemented the above procedure in Python in
a program quartetsampling that samples an alignment randomly to generate many
representative quartet topology replicates for each internal branch in a corresponding focal tree
(https://github.com/fephyfofum/quartetsampling). This procedure has a number of advantages
over NBS for larger datasets. First, alignment columns are not resampled (as in NBS and RBS),
which allows even very sparse alignments to be used. Second, the number of likelihood
calculations that are required is the number of internal branches in the tree multiplied by the
number of replicates per branch multiplied by three possible topologies. Since computation time
scales linearly with the number of taxa, individual replicates are fast, and the computations can be
readily parallelized across processors and furthermore discretized across systems (with results
combined later). This allows QS to be efficiently applied to large alignments beyond the practical limits of NBS and PP. The most extensive computational time was for the Zanne et al. (2014b) 31,749 taxon dataset (see below), which we ran on the Wake Forest University DEAC high-performance cluster using 8 nodes with 16 CPU each. This analysis completed 200 replicates for the full tree in 13 hours. Smaller genome-wide datasets finished 1000 gene-tree replicates on quad-core desktops approximately 12 hours, and conventional multi-gene datasets took only a few minutes to a few hours to run on a standard desktop.

Although the SH-aLRT was by far the fastest method we consider here, the QS was fast enough for large scale analyses. QS can also be applied separately to only a few branches, allowing for more thorough exploration of particular branches of interest. Furthermore, the QS does not require the tree tested to be the maximum likelihood topology, a requirement for SH-aLRT. For our simulated data, we found that performing 200 QS replicates per branch was adequate to achieve low variance in QS score. As would be expected, more replicates per branch should generally be used for larger trees to sample a greater fraction of the total possible quartets.

Furthermore, some branches, especially in large trees, may be entirely unsupported by the alignment due to a lack of sampling overlap among appropriate taxa (i.e., no sites in the alignment contain data from each of the four subsets of taxa; Fig. 1A). Therefore, no phylogenetic information exists to inform the branch (i.e., are “‘indecisive” sensu Steel and Sanderson, 2010), and the QS procedure identifies these branches rather than discarding them or marking them as simply low support.

**Guidelines for interpretation of QS support values**— An important consideration with any measure used to ascertain confidence is the precise interpretation. In order to facilitate accurate
interpretation of the QS scores, we provide a concise visual description of the tests (Fig. 1) and a table describing example scores and their interpretations (Table 1). Particularly notable as shown in Table 1 and in the results, the QS method can not only “support” or “fail to support” a given branch hypothesis, but also can offer “counter-support” for an alternative branch (as in the IC/ICA scores; Salichos et al., 2014; Kobert et al., 2016). Therefore, even “inaccurate” branch hypotheses can offer information in the form strength of the “counter-support” for an alternative quartet topology (i.e., the degree of negativity of the QC score; for examples see Fig. 6).

The QS scores we have described calculate the sensitivity of the resolution of a particular branch to different combinations of taxa sampled around that branch. Each QS replicate calculates whether the four sampled taxa support the resolution of the branch found in the tree over the alternative resolutions. This framework is similar to the interpretation made by those using taxon jackknife analyses for outgroup sensitivity (e.g., Edwards et al., 2005) and the IC score when used with incomplete trees (Kobert et al., 2016). We argue that this interpretation is richer in information than the NBS, and in our simulations the QC score also appears to more conservatively and accurately assign high support values to branches that are present in the true tree (i.e., relatively low false positive rates, at least when the likelihood threshold is small, i.e., in the range of ∼2 used here; Appendix S2). QC scores are particularly helpful in terms of clarifying strength of support for branches with concordant tree frequencies not close to 1 (Appendix S3).

Generation and evaluation of simulated phylogenies— We first tested the method by generating simulated phylogenies under the pure birth (birth = 1) model of evolution with 50, 100, and 500 tips using pxbdsim from the phyx toolkit (Brown et al., 2017). Using these trees, we generated 1000 bp alignments (no indels) under the Jukes-Cantor model with INDELible v.
1.03 (Fletcher and Yang, 2009). Trees were scaled so that the average branch lengths were about 0.2, based on the observation that this generated reasonable trees with most branches recovered correctly from ML analyses. Using the same procedure, we also simulated trees with 500 tips and associated alignments with ten nucleotide partitions, each with 500 sites under the Jukes-Cantor model. We simulated both the full alignment with partitions and a modified randomly resampled sparse alignment to examine the behavior of QS in the presence of missing data (see Appendix S1 for details). These partitioned and sparse alignments had the same qualitative features as the full alignment.

Unlike the NBS method, which generates a set of trees from which branch support is estimated, the QS method requires only a single input topology for which branch support will be measured. We calculated QC, QD, QU, and QF scores for the true underlying tree as well as the ML tree generated by RAxML, but we focus on results for the ML tree. To examine how the number of replicates impacts the QS precision, we conducted simulations varying the number of replicates for randomly drawn branches in the simulated trees (Fig. 2A; Appendix S4). Based on these simulations, we elected to use 200 replicates per branch, since the variation in the QC score was generally low across all tree sizes when this many replicates were performed. We used RAxML and PAUP* to estimate the ML for the three alternative topologies for each QS replicate (using the -f N option and the GTRGAMMA model in RAxML). We also calculated branch-specific QC/QD/QU and taxon-specific QF scores using likelihood differential cutoffs of \( \Delta L = 0 \) (no filtering) and \( \Delta L = 2.0 \), which requires stronger conflicting signal to interpret branches in the input tree as unsupported.
Testing of Empirical Datasets— While the simulations show the general reliability of the method, our primary goal was to use this test to holistically evaluate recent large-scale plant phylogenies, and also to show the utility and versatility of this method in subgroup and even sub-genera analyses. Using published phylogenies and public datasets, we evaluated five recent large-scale phylogenies, including (1) an 103-transcriptome dataset spanning Viridiplantae from Wickett et al. (2014, abbreviated hereafter as WI2014), (2) two large-sparse phylogenies spanning land plants from Hinchliff and Smith (2014b, abbreviated HS2014) and Zanne et al. (2014b, abbreviated ZN2014), and (3) phylogenies spanning Magnoliophyta (angiosperms) with hundreds of genes from Xi et al. (2014a, abbreviated XI2014) and Cannon et al. (2015b, abbreviated CN2015). Additionally, to demonstrate the utility of this method at medium and short time scales, we evaluated two whole transcriptome datasets from the wild tomato clade Solanum sect. Lycopersicon from Pease et al. (2016b, abbreviated PE2016) and carnivorous plants from the order Caryophyllales from Walker et al. (In Press, abbreviated WA2017). Finally, we tested this method on a more typical medium-sized multi-locus dataset from Polypodopsida (ferns) from Pryer et al. (2016b, abbreviated PR2016), such as might appear in many phylogenetic studies of large subgroups. Data for these studies were obtained from datadryad.org and iplant.org (Hinchliff and Smith, 2014a; Matasci et al., 2014; Xi et al., 2014b; Zanne et al., 2014a; Cannon et al., 2015a; Pease et al., 2016a; Pryer et al., 2016a; Walker et al., 2017).

In addition, we analyzed the datasets using 200 individual gene trees for XI2014 and WA2017, and 1000 gene trees for PE2016 and WI2014. For these datasets, quartets are sampled as usual, but only the individual gene sequence alignments are assessed. These phylogenies were all evaluated using a minimum alignment overlap per quartet of 100 bp and a minimum likelihood differential of 2 (i.e., the optimal tree’s log-likelihood must exceed the second-most likely tree by
a value of at least 2). We also calculated the phylogenies with and without partitioning in RAxML, but in all cases the partitioned datasets did not qualitatively differ from the results of the unpartitioned datasets. These data are provided as supplementary data, but are not shown here. We also either re-calculated other measures of branch support or used values from the published studies for comparison to the QS method for each phylogeny, except HS2014 and ZN2014 where the size and sparseness of the datasets prohibited the calculation of other measures of support. For the datasets from CN2015, PR2016, WA2017, and XI2014 100 replicates each of RAxML NBS and SH-test were performed. Additionally, PP scores for PR2016 were calculated using MrBayes (Ronquist and Huelsenbeck, 2003; Ronquist et al., 2012), and IC scores for calculated for Walker et al. (In Press). For PE2016 and WI2014, RAxML NBS, MP-EST, or IC scores were taken from published values.

RESULTS AND DISCUSSION

Analysis of simulated and empirical datasets shows that QS components are consistent, convergent, and complementary— In order to examine the consistency and reliability of the Quartet Sampling method, we first tested QS on a set of simulated phylogenies. As expected, the QC score was generally correlated in a sigmoid fashion with the frequency of concordant trees (Appendix S5). In order to determine the appropriate number of testing replicates and demonstrate convergence, we conducted an analysis of the variance of the QS score by running the QS analysis 100 times on randomly selected branches with different numbers of per-branch replicates for trees with 50, 100, and 500 taxa. This analysis showed that the QC scores converge (with decreasing variance as expected) on a consistent mean value for each branch as the number
of replicates increased (Fig. 2A). Sampling 200 quartets per branch reduced the variance to less than 0.003 in all cases. Similar patterns of convergence were also found for QD values, though with high variance (Appendix S5). We also note that since QS is a branch-specific test (not a tree-wide test), key branches of interest can be tested individually at much higher numbers of replicates without the need to re-test the entire tree.

QC and QD both incorporate the frequencies of discordant quartets in different ways, but for that same reason run the risk of redundancy. However, we observed in the Hinchliff and Smith (2014b) dataset (abbreviated as HS2014) that QD can take on a range of values for QC (Appendix S5). As QC goes to the limits of its range (–1,1), QD values tend to have more extreme values. These extreme values for QD were either due to a lack of discordant trees that limits sampling (when QC approaches 1) or a high amount of one discordant tree (when QC approaches –1). This made QC and QD related, but not strictly correlated, measures. Overall, the mean QC for the HS2014 (0.15; interquartile range (IQR) = [–0.13, 0.46]) and Zanne et al. (2014b) (abbreviated ZN2014; 0.17; IQR = [–0.10, 0.63]) were low compared to the less speciose phylogenies (Fig. 2B; Appendix S5). Applying a minimum log-likelihood differential threshold to small trees tended to push scores toward extremes, resulting in more 0s and 1s (Appendix S2). As expected, given the sparsity of the matrices for ZN2014 and HS2014, the proportion of uninformative quartets was high in both cases (mean QU of 0.65 and 0.85, respectively).

Notably, QS found that 33.4% and 29.8% of branches in HS2014 and ZN2014, respectively, had QC values less than –0.05. This meant that about a third of the branches in these consensus phylogenies reported not just “low support” for the given branch, went further to report “counter-support” for one of the two alternative topological arrangements at that branch (see Table 1). Many empirical cases have been documented where one anomalous lineage can distort
not only its own phylogenetic placement, but can exert a wide-ranging set of distortions on the
topology (e.g., Fig. 8A, among many others). Resampling methods that infer a complete set of
tips (like NBS) will produce replicates that may also be distorted and thus firmly support these
relationships rather than highlight this discordance. Therefore, as pointed out by many other
quartet methods (Mirarab et al., 2014; Sayyari and Mirarab, 2016), the use of quartets breaks
down the problem in more manageable units that cannot be distorted by these complex lineage
sorting effects.

QS indicates strong relationships within, but not among, major land plant lineages— While
simulations and general characterizations of the QS method are necessary to demonstrate its
effectiveness, the primary goal of this study was to use QS to reanalyze and compare several
recent speciose and phylogenomic datasets to address ongoing debates of phylogenetic
relationships in the plant tree of life. We used QS methods to evaluate two of the most speciose
phylogenies of land plants presently assembled by Hinchliff and Smith (2014b) and Zanne et al.
(2014b), and one of the most comprehensive phylogenies of Viridiplantae from Wickett et al.
(2014) (abbreviated as WI2014). HS2014 spanned 13,093 taxa with a total alignment of 148,143
sites (Fig. 3), but extremely sparse overall coverage (96.4% missing characters). The alignment
used for ZN2014 (Fig. 4) contained 31,749 taxa and an alignment length of 12,632 bp, but also
with sparse coverage (82.3% missing characters). However, the QS method accommodates large,
sparse matrices, by evaluating alignment overlap only for each replicate quartet without requiring
complete occupancy for any partition of the alignment. Finally, the WI2014 (Fig. 5) contains only
80 taxa, but, as the pilot study for the 1000 Plants Project (1KP), spans the deepest time scale of
these three (including Chlorophyta) and had a highly occupied alignment of 290,718 bp with only
10.7% missing data.

Most major plant groups showed strong support in HS2014 and ZN2014, including angiosperms (QC=0.68 and 0.75, respectively), Anthocerotophyta (hornworts; QC=0.54,0.92), Acrogynnospermae (gymnosperms; QC=0.37,0.32), and Polypodopsida (ferns; QC=0.23,0.46). Marchantiophyta (liverworts), rosids, and asterids all showed low support in HC2014 (QC < 0.2). Monocots (both including and excluding Acorus) had poor support in both HS2014 and ZN2014 (–0.05 < QC < 0.05). In WI2014, all the major groups (i.e., labeled internal branches in Fig. 5) showed high support. In contrast to the generally high support for these major established groups, we frequently found low support on the “backbone” relationships among these groups, in a manner consistent with most previous phylogenies of land plants. The relationships among Marchantiophyta, Bryophyta (mosses), Anthocerotophyta, lycophytes, and “euphyllophytes” (i.e., ferns and seed-bearing plants) has also been under substantial ongoing debate (Shaw et al., 2011). HS2014 places mosses as sister to the remaining land plants, followed successively by liverworts, hornworts, and lycophytes. However, HS2014 indicates QC=–0.04 for a branch defining liverworts + all other land plants the exclusion of mosses (Figs. 3B, 6A), indicating that the most common quartet branch among the replicates was not the branch shown in the tree. By contrast WI2014 shows strong support (QC=0.67) for the branch leading to a moss+liverwort common ancestor (Fig. 5). While ZN2014 contains only embryophytes, it shows weak support (QC=0.15) for the branch separating moss+liverwort from the rest of land plants. WI2014 has hornworts (represented by two Nothoceros) as sister to a clade of mosses and liverworts+tracheophytes (as in Renzaglia et al., 2000), with high support for all relationships. However, note that WI2014 showed substantial counter-support against a common mosses+liverworts+tracheophyte ancestor to the exclusion of hornworts. Coalescent analyses in
WI2014 also support this finding of a monophyletic Bryophyta+Marchantiophyta clade.

Therefore, while the topology of HS2014 was consistent with the order of many previous phylogenies (Nickrent et al., 2000; Qiu et al., 2006; Chang and Graham, 2011), QS supported the alternative configuration of mosses and liverworts as sister groups (as in WI2014 and ZN2014; see also Zhong et al., 2013a).

In all three datasets, the monophyly of vascular plants was strongly maintained, despite the inclusion of Selaginella that have unusual GC content (Banks et al., 2011). However, the branch leading to Selaginella often was accompanied by a higher QD value, which could be a result of the biased composition or indication of a secondary evolutionary history. The QF value for Selaginella in all cases was moderate, indicating that overall it did not more often produce discordant topologies compared to other lineages. We also observed substantial discordance and counter-support for relationships tested among various bryophyte groups and key taxa in HS2014, possibly indicative of substantially under-appreciated hybridization among mosses (Nylander et al., 2004). The collective results indicated a (perhaps complicated) monophyly of liverworts and mosses, inconsistent placement of hornworts, and strong support for tracheophytes.

QS analysis of ferns confirms close relationship to seed plants and shows QS is effective on smaller datasets— Another notable similarity of both HS2014 and ZN2014 was the poor score at the base of “euphyllophytes” that placed ferns in a clade with seed plants. QC scores on this branch were near zero in HS2104 and ZN2014 (0.02 and −0.06, respectively) and relatively low for the WI2014 tree (QC=0.33); QD values were also relatively large (0.44–1). Within ferns the arrangement of major clades in ZN2014 (Fig. 4E) was mostly consistent with the recently published phylogeny by The Pteridophyte Phylogeny Group (PPG I, 2016). Those clades whose
relationships were counter-supported (Marratiales, Salviniales, Hymenophyllales) were discordant with the PPG I consensus and other recent phylogenies (Pryer et al., 2004; Testo and Sundue, 2016) demonstrating the utility of QS in highlighting suspect relationships. Some key areas of known high uncertainty (e.g., Saccoloma and Lindsaea) were highlighted with low and negative QC scores. Additionally, the debated position of Equisetum was also reflected by low scores.

While QS was designed for large datasets, we also found that QS can perform well on smaller multi-gene datasets conventionally used for systematics studies. Using the phylogeny and data from Pryer et al. (2016b), we reanalyzed the 5778 bp alignment using QS, NBS, SH, and PP scores (Fig. 7). The QS scores were more conservative, but confirmed the conclusions of Pryer et al. (2016b) regarding the monophyly of maidenhair ferns (Adiantum) and its placement in a clade with the Vittarioids. In terms of QC/QD/QU scores, ZN2014 also supported the monophyly of Adiantum (0./1.0/0.83; 14 taxa), the Vittarioids (0.20/1/0.71; 8 taxa), and their common ancestor (0.22/1/0.87). We also found that among different branches in PR2016 with fairly similar NBS and PP values, QS offers a wide range of support values (including negative-QC counter-support).

Gymnosperms are monophyletic under QS and support Ginkgo in a clade with cycads— Another question that has attracted substantial debate is the relationships among the major gymnosperm lineages and angiosperms. Under QS evaluation, HS2014, ZN2014, XI2014, and WI2014 all indicate strong support for monophyly of recognized gymnosperm lineages. However, the relationships among cone-bearing lineages differ among these four phylogenies. ZN2014 and WI2014 place Ginkgo in a clade with cycads (consistent with Qiu et al., 2006; Bowe et al., 2000;
Lee et al., 2011; Xi et al., 2013; Fig. 6B). While the HS2014 topology places cycads as sister to the remaining gymnosperms (i.e., not monophyletic with *Ginkgo*), the QS evaluation counter-supports this relationship. Therefore, even though HS2014/WI2014 and ZN2014 had different topological arrangements of these taxa, the QS analyses of these datasets indicates a consistent message of a *Ginkgo+cycads* clade separate from the rest of gymnosperms.

This pattern of disagreement in topology but consistent QS interpretation was observed again in the placement of Gnetales relative to the conifer linages (Fig. 6C). ZN2014 indicates Gnetales in a clade with a monophyletic Pinales (consistent with Lee et al. 2011) whereas HS2014 and WI2014 show a Gnetales+Pinaceae ancestor distinct from other conifers (i.e., the “Gnepine” hypothesis; Bowe et al., 2000; Xi et al., 2013). However, again the HS2014 and WI2014 QS scores offer counter-support against the “Gnepine” relationship (QC=-0.19 and -0.67, respectively) and show that one alternative was strongly preferred (QD=0.5 and 0.1). Here again, in evaluating different competing topologies the QS method can offer consistent interpretations that in this case indicates the monophyly of Pinales, but perhaps also offer some (albeit weak) evidence that warrants further examination of possible gene flow between Gnetales and Pinales.

Finally, among the non-Pinaceae conifers, we observed conflicting patterns of poor backbone support for monophyletic clades. In both ZN2014 and HS2014, *Amentotaxus* and *Torreya* form a clade with the rest of Taxaceae to the exclusion of *Cephalotaxus*. Relationships between these three Taxaceae lineages are both strongly counter-supported by QS. The high values of QD (0.56–0.8) offer indication of possible introgression among these lineages. Therefore, again QS scores highlight a part of the phylogenetic tree that may be mis-modeled by these consensus phylogenies.
Quartet Sampling indicates biased conflict among the ANA grade angiosperms— Few issues in angiosperm evolution have garnered more recent debate than the relationship among the so-called “ANA grade” angiosperms, which include Amborella, Nymphaeales, and Austrobaileyales. Of the four datasets where this phylogenetic relationship was testable, only WI2014 inferred the “Amborella-first” hypothesis (Figs. 5, 6D; Qiu et al., 1999; Stefanović et al., 2004; Qiu et al., 2006; Moore et al., 2007; Soltis et al., 2011; The Amborella Genome Project, 2013). ZN2014 shows a “Nymphaeales-first” relationship (Fig. 4B), and HS2014 and XI2014 indicate “Amborella+Nymphaeales”-first (Fig. 3B; Appendix S6). Regardless of the resolution inferred in the consensus phylogenies from each dataset, the branches surrounding the ANA-grade were all counter-supported (QC < 0) and biased in their discordance (QD > 0.8; Fig. 6D). ZN2014 offers weak support for Amborella+Nymphaeales, while XI2014 counter-supports this relationship. In HS2014, WI2014, and ZN2014, the scores on the branch that placed Austrobaileyales as the closest relative group to the remainder of angiosperms were strongly QC-counter-supported and QD-biased.

Two questions surround the relationships of the ANA grade angiosperms. First, what was the relationship among these lineages? Second, are the longstanding disagreements in inference of these relationships the result of genuine biological conflict (i.e., introgression, horizontal transfer, etc.), limitations in the data, or a methodological artifact arising from the depth of this branch, the monotypic status of Amborella, and/or the rapidity of the angiosperm radiation? On the first question, QS lacks support for “Nymphaeales-first”. Whether there more support exists for Amborella+Nymphaeales or “Amborella-first” (as found also by The Amborella Genome Project, 2013) is unclear. The results we present suggest that both relationships have support in the data, depending on changes in the dataset composition (as also found by Wickett et al. 2014).
On the second question, however, the strong QD values indicate a biased conflict that suggests a secondary evolutionary history conflicting with the primary tree in a way that has so far evaded comprehensive characterization. Demonstrations that bryophyte mitochondrial sequences are present in *Amborella* (Rice et al., 2013; Taylor et al., 2015) indicate that these lineages have experienced ancient introgressive events that might create competing evolutionary narratives in the manner indicated by the QS results shown here (and see also Shen et al., 2017). Given the intensity of effort to address these relationships without a broad community consensus and the specific evidence of long-range introgression in *Amborella*, a greater understanding of ANA-grade evolution likely lies in an examination of complex evolutionary histories rather than in a continuation of the debate over appropriate sampling or models.

**QS supports a magnoliid+eudicot ancestor and raises questions about the placement of the Chloranthaceae and Ceratophyllaceae**— The timings and order of the relationships among the three “core angiosperm” lineages (eudicots, monocots, and magnoliids) represent the evolution of three clades that have transformed the biosphere. ZN2014, WI2014, and XI2014 all indicate the existence of a magnoliid+eudicot clade (Figs. 4B, 5 Fig. 6E, Appendix S6; also in Qiu et al., 2006; Burleigh et al., 2009; Lee et al., 2011). An alternative monocot+eudicot clade appears in HS2014 (Fig. 3B), which has been suggested by many other past phylogenies (Jansen et al., 2007; Moore et al., 2007; Qiu et al., 2010; Soltis et al., 2011). However, in HS2014 the QS scores counter-support both eudicot+monocots and magnoliids/monocot+eudicot clades. Here again, despite disagreement among the topologies of the three large-scale phylogenies, the supporting and counter-supporting QS scores indicate a common evolutionary history.

In addition to the three major core angiosperm groups, the Chloranthaceae have frequently
been placed in a clade with magnoliids (as in Jansen et al. 2007; Moore et al. 2007; Soltis et al. 2011; Group et al. 2016, and Fig. 3 in Wickett et al. 2014). However, both HS2014 and ZN2014 place Chloranthaceae in a clade with magnoliids, monocot, and eudicots, while WI2014 places it in a clade with just eudicots (excluding monocots and magnoliids) that is, however, counter-supported by the sampled quartets (QC=–0.54, QD=0.95).

*Ceratophyllum*, often placed as the most closely related group to eudicots (Jansen et al., 2007; Soltis et al., 2011), in HS2014 appears in a clade with the Chloranthaceae among the prior to monocot/magnoliid/eudicot splits. However, the branch placing Chloranthaceae+*Ceratophyllum* lineage in a clade with magnoliids+monocots+eudicots was counter-supported (Fig. 3B) with a strongly biased QD (0.81). ZN2014 supports *Ceratophyllum* as separate from the core angiosperms. Given that this question is inextricably linked with the relationships of the three core angiosperm groups, a consensus was not reached by this information alone. However, this evidence reopens the question of the relationship of Chloranthaceae and Ceratophyllaceae, both to each other and to the ANA grade angiosperm lineages (see discussion in Eklund et al., 2004).

**Relatively consistent topology but poor support found within monocots**— Within monocots, *Acorus* was weakly but consistently supported as a separate clade from the rest of monocots in HS2014, ZN2014, and WI2014 (Figs. 3B, 4B, 5). In general, the arrangement of monocot orders in both HS2014 (Fig. 3C) and ZN2014 (Fig. 4C) agreed with recent consensus phylogenies (Givnish et al., 2010; Soltis et al., 2011; Barrett et al., 2015; Givnish et al., 2016; McKain et al., 2016).

One exception was the placement of Liliales, which has varied in consensus trees and, here, either was sister to commelinids with weak support (HS2014) or was sister to Asparagales.
The commelinid orders were resolved inconsistently between studies. The commelinid stem branch was counter-supported in both ZN2014 and HS2014 with high QD values. In HS2014, Arecales were separate from the rest of the commelinids, but the ancestral commelinid branches both including and excluding Arecales were counter supported. In ZN2014, Arecales was supported in a clade with Zingiberales and Commelinales, but the ancestral branch for these three orders has a QD=0.79. This arrangement was consistent with previous phylogenies (e.g., Givnish et al. (2010)) that have poor support to these relationships. From the QS results, we would cautiously infer that (1) the relationships among the commelinids are still unknown, (2) there may be uncharacterized secondary evolutionary history distorting the phylogenetic placement of these groups, and (3) likely the variable data from both Liliales and Arecales together have a joint effect that is causing inconsistency in the phylogenetic inference. Finally, in Poaceae, Quartet Sampling made clear the well-characterized discordance and complex relationships (e.g., Washburn et al., 2015; McKain et al., 2016). The “BOP” clade (Bambusoideae, Oryzeae, Pooideae) was counter-supported in HS2014 and both paraphyletic and counter-supported in ZN2014. Within the “PACMAD” clade, many of the named subgroupings were counter-supported with negative QC values. There were particularly high QU values in both HS2014 and ZN2014 for this clade indicating not only complex support but also low information. Therefore, even if someone were unfamiliar with the controversies surrounding monocot phylogenetics (as a whole and within groups), the QS evaluation framework clearly highlights this group as one with consistently high conflict and counter-support values for the consensus phylogenies in HS2014 and ZN2014.
Non-rosid/asterid eudicot lineages show substantial disagreement and rogue taxa— Within eudicots, Ranunculales were supported as a separate clade from the rest of eudicots by WI2014, HS2014, XI2014, and CN2015, or among the non-rosid/asterid eudicots in ZN2014 (Figs. 3B, 4B, 5; Appendix S6, Appendix S7). Notably in both ZN2015 and HS2014, *Nelumbo* (sacred lotus) was placed sister to a highly supported Proteales clade, but support for the *Nelumbo*+Proteales clade was strongly counter-supported by the QC and QD scores. This again demonstrates a situation where QS effectively highlights this lineage’s conflicted relationship with the Proteales in a way that raises questions about a possible genuine biological (rather than artifactual) cause of phylogenetic discordance.

*Vitis vinifera* (common grape vine) has also had variable inferred positions throughout eudicots. WI2014, XI2014, and CN2015 all place *Vitis* as a “superrosid” and most closely related clade to the rosids, but with widely ranging support and skew (QC=–0.13 to 0.63, QD=0.56 to 1.0) consistent with the conflict noted by Cannon et al. (2015b) (abbreviated CN2015; Appendix S7). HS2014 (which sampled 64 Vitales) supported a Dilleniales+Gunnerales+Vitales clade, while ZN2014 places Vitales+*Tetrastigma* among the rosids in a Vitales+malvid clade to the exclusion of fabids. In the case of the grape vine, the high support in WI2014 (analyzed by gene trees) accompanied by a high QD value (1.0) gives strong indication that while Vitales may belong among the superrosids, an investigation of a secondary evolution history for grapes might be warranted.

For the relationships among the superasterid groups (Caryophyllales, Berberidopsidales, Santalales, and asterids), a common pattern was found in HS2014, ZN2015, WI2014, and XI2014 of near-zero QC values (–0.03 to 0.08), modest QD values (0–0.03), and high QU values (0.49–0.86). This led to a consensus QS interpretation of low phylogenetic information, likely as a
result of the rapid radiation of these lineages. Generally, these phylogenies tended to support
weakly the controversial placement of Caryophyllales as most closely related to the eudicot
ancestor.

QS shows strong support for the consensus rosid phylogeny, but substantial discordance
both within and among asterid clades— HS2014 (Fig. 3D) find similar resolutions among rosid
lineages (Wang et al., 2009; Moore et al., 2010; Soltis et al., 2011; Zhao et al., 2016) (the lone
exception was the Sapindales+Huertales clade). Among the branches leading to major groups, all
were supported and most of the backbone branches were either supported or only weakly
negative. The QS scores also correctly identified a poorly supported relationship in HS2014
between Cynomorium and Cucurbitales (QC=–0.31). Cynomorium, a non-photosynthetic
parasitic plant with unusual morphology, has been placed tenuously and variably in groups as
diverse as Rosales (Zhang et al., 2009) and Saxifragales (Nickrent et al., 2005), so its poor score
here was expected and confirms the QS method’s ability to detect poorly placed species. This
“rogue” status was corroborated by the below-average QF score of QF=0.18 (mean 0.21 for
HS2014). This means that for quartets that include Cynomorium as a randomly sampled taxon,
only 18% produced a quartet topology concordant with the HS2014 tree.

In contrast to the analysis of the rosid phylogeny from HS2014, the asterid phylogeny from
ZN2014 (Fig. 4D) appears to have substantial discordance and disagreement with most published
phylogenies (Soltis et al., 2011; Beaulieu et al., 2013; Refulio-Rodriguez and Olmstead, 2014).
ZN2014 QS scores supported the unusual hypothesis of a common Ericales+Cornales ancestor,
weakly support the campanulid clade, and counter-support a common lamiid ancestor. The
arrangement of families within Asterales either roughly conforms to Soltis et al. (2011) and
Beaulieu et al. (2013), or counter-supports branches (QC<0) that do not agree with these consensus phylogenies. However, most of the branches that define the relationships among asterid orders in ZN2014 were counter-supported by the data, though most have QC and QD values close to zero. This indicates a scenario of a rapid radiation rather than hybridization (though these are not mutually exclusive). The lamiid groups show strong support for the monophyly of the Lamianae core orders, though in an order different from Stull et al. (2015). Among the asterids there were also several notably low QF scores for singular genera like Oncotheca, Vahlia, and Pentaphragma (all QF<0.16). Thus, the QS method in asterids does not necessarily add clarity to this group, but perhaps more concretely puts into focus the combinations of discordance and low information occurring in this group.

QS of shallow-timescale phylotranscriptomic datasets— So far, we have demonstrated the utility of quartet sampling on large, sparse alignments which are often computationally intractable with other support measures. We have also shown, in the case of WI2014, that a relatively large and full occupied matrix from deep-timescale transcriptomic data can also be evaluated by QS. However, the QS method can be used to rapidly evaluate phylogenetic support on genome-wide datasets with little missing data for shorter evolutionary timescales. We tested the QS method on two phylotranscriptomic datasets for the wild and domesticated tomato clade Solanum sect. Lycopersicon (Fig. 8A; Pease et al., 2016b) and carnivorous plants spanning the Caryophyllales (Fig. 8B; Walker et al., In Press).

The Solanum phylogeny from Pease et al. (2016b) was inferred from coding sequence alignment of 33,105,168 nucleotide sites for 30 populations spanning all 13 wild and domesticated tomato species, and two outgroup species. As described in Pease et al. 2016b, this
dataset contains a high level of phylogenetic discordance, but had a consensus phylogeny with 100% NBS support at all but two branches (occurring within a large species complex). However, gene tree analysis of this group clearly showed evidence of massive phylogenetic discordance. When we applied QS to this phylogeny using the entire alignment, scores for many branches were also perfect (i.e., 1/0/0; Table 1). However, several of the other branches in the “Peruvianum group” species complex had lower QS scores in the full alignment (Fig. 8A). When gene trees were used (both a quartet of taxa and a gene alignment were randomly chosen for 1000 QS replicates), all branches had QC<1 in a manner consistent with the gene tree discordance found previously in this clade. We also observed the presence of high QD values within the major subgroups reported for this clade, while nodes defining these groups showed high QC and low QD values. This accurately captures the results found by Pease et al. (2016b) in terms of the low discordance between groups versus high discordance within the major groups.

Most notably, the tree shown in Fig. 8A includes S. huaylasense accession LA1360. This accession has been known (both from this and other datasets) to mostly likely be a hybrid between populations from the green-fruited and red-fruited lineages (essentially those accessions above and below LA1360, respectively, in Fig. 8A). Thus, the inclusion of this putative hybrid lineage distorted the phylogeny as tree inference methods tried to cope with inherited and introgressed alleles from two separate groups to place this accession in a consensus location on the tree. While NBS scores were high for the branches surrounding the placement of LA1360, QS showed negative QC scores and high QD scores (QD=1 for full alignment). This supports the presence of the alternative phylogenetic history that has been previously corroborated by other studies (see additional discussion in the Supplementary Results of Pease et al. 2016b). These data show clearly that QS was able to distinguish between consistently supported relationships and
branches known to have conflict due to introgression (whereas NBS does not).

An analysis of transcriptomes of carnivorous plants from Caryophyllales (Fig. 8B; Walker et al. In Press) showed QC scores consistent with IC scores. The QS scores for the ancestor of “core Caryophyllales” were near zero and the high QD value when gene trees were used (QD=0.68) supported the hypothesis of Walker et al. (In Press) that introgressive gene flow may have occurred among these lineages. The evidence for placing *Drosophyllum* among the carnivorous Caryophyllales has always been tenuous. The QS analysis here showed not only a low QF value (QF=0.76, average for WA2017 was 0.89) for this taxon, but also low-QC/high-QD values for the two branches that form the clade with *Ancistrocladus* and *Nepenthes*. As with the tomato example above, this example demonstrates how QS scores can highlight an entire region that may be distorted by the inclusion of a taxon with a strong potential for a secondary evolutionary history (i.e., possible introgression).

**CONCLUSION**

In this study, our goal was to reanalyze several key and long-contested conflicts in the plant tree of life by embracing the apparently complex relationships in the plant tree of life. The Quartet Sampling method that we have described and demonstrated here synthesizes a blend of phylogenetic and phylogenomic methods to provide a coherent framework for distinguishing several causes of low phylogenetic branch support. QS also provides a tractable means to analyze sparse datasets with tens of thousands of taxa but poor sequence overlap. Bootstrap and posterior probability values have been shown to exhibit irregular behavior or to report uniformly high confidence scores for large-scale datasets, despite substantial underlying conflict. Even when
conventional support measures do report low support, the specific cause of the low support is not indicated, leaving it open to broad interpretation and speculation. Despite the decades of documented concerns about the limitations of these support measures, bootstrap proportions remain the dominant support measure in systematics. Our results, as well as those reported in many other studies cited here, demonstrate the need for continuing discussion about the nature of phylogenetic disagreement itself and how best to quantify this complexity. While much testing and refinement of the QS method is undoubtedly left to be done, we find it provides a key function that has been missing from other support measures, namely the ability to distinguish among different causes of low support that commonly occur in modern molecular phylogenies. The application of the QS method to the plant tree of life has also yielded some intriguing results. The comparison of several recent consensus phylogenies showed that they differ at several key, disputed relationships. However, when QS scores were compared, only one topology was supported with the alternatives showing weak support, conflicted support, or even counter-support (e.g., *Ginkgo* placed in a clade with cycads, *Gnetales* in a clade with a common conifer ancestor, monocots separate from a magnoliid+eudicot clade, and *Caryophyllales* sister to asterids). QS scores for some other relationships indicated the existence of multiple conflicting but supported evolutionary histories (e.g., the placement of *Amborella*, possible widespread gene flow in the monocots, and notoriously difficult-to-place genera like *Haplomitrium*, *Nelumbo*, and *Cynomorium*). The artist Man Ray once remarked that “We have never attained the infinite variety and contradictions that exist in nature.” Overall, the picture painted by QS is one of substantial contradiction, but this conflict can be a richly informative (not just confounding) illustration of the interwoven evolutionary histories contained within the plant tree of life.
ACKNOWLEDGEMENTS

The authors thank Ya Yang, Caroline Parins-Fukuchi, and Kathy Kron for helpful discussions, and Luke Harmon, Eric Roalson, and Matt Pennell for valuable feedback on drafts. Computations were performed on the Wake Forest University DEAC Cluster, a centrally managed resource with support provided in part by the University.

FUNDING

SAS and JWB were supported by National Science Foundation Assembling, Visualizing, and Analyzing the Tree of Life Grant 1208809.

LITERATURE CITED


 genomic compartments: extant gymnosperms are monophyletic and Gnetales’ closest relatives
 are conifers. *Proceedings of the National Academy of Sciences of the United States of America*
 97: 4092–4097.

Brown, J. M. and A. R. Lemmon. 2007. The importance of data partitioning and the utility of


*Bioinformatics*.


 et al. 2015a. Data from: Multiple polyploidy events in the early radiation of nodulating and

 et al. 2015b. Multiple polyploidy events in the early radiation of nodulating and nonnodulating

Chang, Y. and S. W. Graham. 2011. Inferring the higher-order phylogeny of mosses (Bryophyta)
 and relatives using a large, multigene plastid data set. *American Journal of Botany* 98:
 839–849.


(Adiantum) are indeed monophyletic and sister to shoestring ferns, vittarioids (Pteridaceae).

Systematic Botany 41: 17–23.


Simmons, M. P. 2016. Mutually exclusive phylogenomic inferences at the root of the angiosperms: *Amborella* is supported as sister and Observed Variability is biased. *Cladistics*.


Table 1: Quartet Sampling (QS) score interpretation. See text for details.

<table>
<thead>
<tr>
<th>Example QS Score (QC/QD/QU)</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0/-/0.0</td>
<td><strong>Full support:</strong> All sampled quartet replicates support the focal branch (QC=1) with no uncertainty when likelihood cutoffs are used (QU=0).</td>
</tr>
<tr>
<td>0.5/0.02/0.03</td>
<td><strong>Strong support:</strong> A strong majority of quartets support the focal branch (QC=0.5) and the low skew in discordant frequencies (QD≈0) indicate no alternative history is favored.</td>
</tr>
<tr>
<td>0.7/0.9/0.03</td>
<td><strong>Strong support with discordant skew:</strong> A strong majority of quartets support the focal branch (QC=0.7), but the skew in discordance (QD=0.9) indicates the possible presence of a supported secondary evolutionary history.</td>
</tr>
<tr>
<td>0.05/0.04/0.03</td>
<td><strong>Weak support:</strong> Only a weak majority of quartets support the focal branch (QC=0.05), and the frequency of all three possible topologies are similar (QD≈0).</td>
</tr>
<tr>
<td>0.1/0.9/0.03</td>
<td><strong>Weak support with discordant skew:</strong> Only a weak majority of quartets support the focal branch (QC=0.1), and the skew in discordance (QD=0.9) indicates the possible presence of a supported secondary evolutionary history.</td>
</tr>
<tr>
<td>-0.5/0.9/0.07</td>
<td><strong>Counter-support:</strong> A strong majority of quartets support one of the alternative discordant quartet arrangements history (QC&lt;0; QD expected to be high).</td>
</tr>
<tr>
<td>1/0.03/0.95</td>
<td><strong>Poorly informed:</strong> Despite supportive QC/QD values, 95% of quartets failed the likelihood cutoff (QU=0.95), likely indicating few informative sites.</td>
</tr>
<tr>
<td>0.0/0.0/0.0</td>
<td><strong>Perfectly conflicted:</strong> The (unlikely) case where the frequency of all three possible trees was equal with no uncertainty, indicating a rapid radiation or highly complex conflict.</td>
</tr>
</tbody>
</table>

Notes: * QC = Quartet Concordance; QD = Quartet Differential; QU = Quartet Uncertainty.
**FIGURE LEGENDS**

**Fig. 1.** Description of the Quartet Sampling method. (A) The focal branch “b” divides the phylogeny into four subclades \( \{S_1, S_2, S_3, S_4\} \) from which tips (A–J) are sampled. Two replicates with different sampled tips for the given branch are shown with the three possible unrooted topologies (one concordant and two discordant). (B) Each internal branch is labeled with a set of three scores (QC/QD/QU), which offer different, but complementary, information. Terminal branches are evaluated by the QF score, which reports the frequency of a taxon generating concordant topologies. (See Materials and Methods for full details and Supplementary Methods for a technical description.)

**Fig. 2.** Results of Simulation Testing of the Quartet Sampling Method. (A) QC values converge on a central value with increasing numbers of replicates from randomly selected branches from simulated trees with 50, 100, and 500 taxa. (B) Distributions of QC, QU, and QF values for HS2014 (black), ZN2014 (dotted black), and XI2014/CN2015/PR2016/WA2017 (similar distributions; gray solid). (C) Mean QC values (diamond) with 5%ile to 95%ile (whiskers) for branches in HS2015 binned by the number of subtending taxa (i.e., moving root-ward in the tree left-to-right). Overall mean is shown with horizontal dotted line.

**Fig. 3.** Phylogeny from Hinchliff and Smith (2014b). (A) Full phylogeny with heat map coloration of branches by QC scores for internal branches: dark green (QC>0.2), light green (0.2≤QC>0), light orange (0≤QC≤–0.05), or dark orange (QC<–0.05). (B) QC/QD/QU scores (200 replicates of full alignment) for major plant groups and key orders within angiosperms. QC/QD/QU scores after group names are for the ancestral branch (i.e., the “stem” branch), and a single QF score is shown for monotypic tips. Major subgroups groups are highlighted with vertical labels. (C) QS scores for monocots (excluding *Acorus*). (D,E,F) QS scores for rosids, Bryophyta, and gymnosperms. Abbreviations: Acro, Acrogymnospermae; ANA, ANA grade; Aru, Arundinoideae; Bry, Bryophyta, Chl, Chloridoideae; Dan, Danthonioideae; Mar, Marchantiophyta; Poly, Polypodopsida.

**Fig. 4.** Phylogeny from Zanne et al. (2014b). (A) Full phylogeny with heat map coloration of branches by QC scores.
for internal branches using same color scheme as (Fig. 3). (B) QC/QD/QU scores (200 replicates of full alignment) for major plant groups and key orders within angiosperms, using same color scheme as (Fig. 3). (C) QS scores shown for monocots (except Acorus). (D) QS scores for asterids. (E) QS scores for fern lineages and (F) QS scores for gymnosperm lineages respectively. Abbreviations: Alseu, Alseuosmiaceae; ANA, ANA grade angiosperms; Argo, Argophyllaceae; Aster, Asteraceae; Bory, Boryaceae; Caly, Calycanthaceae; Eriach, Eriachneae; Good, Goodeniaceae; gym, gymnosperms; Hypox, Hypoxidaceae; Isach, Isachneae; Phell, Phellinaceae; Poly, Polypodopsida.

Fig. 5. Maximum likelihood phylogeny spanning Viridiplantae from Fig. 2 in Wickett et al. (2014) with QC/QD/QU scores for 200 replicates of the full alignment. Nodes are colored according to QC score using same color scheme as (Fig. 3). Bootstrap values (italicized in square brackets) from Wickett et al. (2014) are shown for comparison. Missing QS or bootstrap values indicate a perfect score. The three taxa with the lowest QF values are highlighted. Species names have been excluded or abbreviated in the case where two congeners are included.

Fig. 6. Key phylogenetic disagreements with QC scores using compared across various datasets. Branches for HS2014 and ZN2014 were resampled with 10000 replicates. Branches for WI2014 and XI2014 were exhaustively sampled (>1000 replicates). Highlighting on QC values follows the same colors as Fig. 3. “Conifers-II” refers to a hypothesized clade comprising the non-Pinales orders in Pinidae. Abbreviations: Gnet, Gnetidae; Pin, Pinidae.

Fig. 7. Phylogeny of Pteridaceae ferns from Pryer et al. (2016b) with QC/QD/QU scores for 200 replicates of the full alignment. Nodes are colored according to QC score using same color scheme as (Fig. 3). Bootstrap/SH-test/posterior probability values (italicized in square brackets) are shown for comparison. Omitted values indicate a perfect score. The three taxa with the lowest QF values are highlighted. Abbreviations: Pityro, Pityrogramma.

Fig. 8. QS scores for phylogenies from whole-transcriptome data. Omitted values indicate a perfect score. Nodes are colored according to QC score using same color scheme as (Fig. 3). (A) Phylogeny of Solanum sect. Lycopersicon.
from Pease et al. (2016b) Bootstrap values (italicized in square brackets) are shown for comparison. (B) Phylogeny of Caryophyllales from Walker et al. (In Press) IC scores (light grey) are shown for comparison (all bootstrap and SH-test scores were 100). The three taxa with the lowest QF values are highlighted.
APPENDICES

Appendix S1. Supplementary Methods providing a technical description of the QS method

Appendix S2. Comparison of QC and bootstrap ICA (information criterion-all; Salichos, et al. 2014) scores on trees reconstructed from 100 simulated datasets with 50 taxa with 1,000 base pairs under a Jukes-Cantor model of evolution. Blue circles represent branches in the true tree, with the size of the circle proportional to the log of the number of substitutions. Red triangles represent branches not in the true tree.

Appendix S3. Comparison of the rapid bootstrap and quartet sampling on the ML/PP consensus tree. For each branch, the RBS, QS (raw concordant frequency (Freq1), QC score), SH, and PP scores are presented (clockwise from top left in each legend). Black dots identify clades that are not in the true tree.

Appendix S4. Shows the consistency of the frequency of concordant quartets ($f_1$), QC, and QD toward a central value with increasing number of per-branch replicates for a randomly selected branch. Trees with 50 taxa (left), 100 taxa (center), and 500 taxa (right) are shown. Boxes show median ± IQR. Whiskers show 5th–95th percentile, with values outside this range shown as circle points.

Appendix S5. Histograms (top row) showing the distributions of QC (left), QU (middle), and QF (right) values for the HS2014 dataset (green), ZN2014 (black), and smaller dataset (XI2014, CN2015, PR2016, WA2017) with similar distributions (orange). Scatter plots (bottom row) showing the close (but non-linear) relationship between QC and raw concordant quartet frequency ($f_1$; left), bounded but otherwise uncorrelated relationships between QC and QD (middle), and QC and QU (right). See main text for dataset abbreviations.

Appendix S6. Phylogeny of angiosperms from Xi et al. (2014a) with QC/QD/QU scores for 200 replicates of the full alignment and for 200 replicates from individual gene trees (in parentheses). Nodes are colored according to QC score using same color scheme as (Fig. 3). MrBayes PP/RAxML NBS values (italicized in square brackets) from Xi
et al. (2013). are shown for comparison. Perfect scores for any given test are omitted or shown as ‘*’ indicates bootstrap of 100, while ‘-’ indicates a missing value. The three taxa with the lowest QF values are highlighted.

Appendix S7. Phylogeny from Cannon et al. (2015b) with QC/QD/QU scores for 200 replicates of the full alignment. Nodes are colored according to QC score using same color scheme as (Fig. 3). Bootstrap values (italicized in square brackets) are shown for comparison. Perfect scores for any given test are omitted or shown as ‘*’ indicates bootstrap of 100, while “-” indicates a missing value. The three taxa with the lowest QF values are highlighted.

Appendix S8. Relationship between QC and frequencies of the three possible alternative quartet topologies from QS runs on simulated data. Points represent branches in the trees, with the “test topology” axis identifying the frequency at which the topology consistent with the tree was recovered across all QS replicates for that branch, and the “alt n” axes identifying the frequencies of the two alternative (conflicting) topologies.
A

Replicate 1
A b H
F J

Replicate 2
B b H
G J

Concordant quartet topology
Discordant topology #1
Discordant topology #2

B

Quartet Sampling
Internal Node Scores = \(0.52/0.04/0.01\)

Quartet Concordance (QC)
How often is the concordant quartet inferred over both discordant quartets?

Quartet Differential (QD)
Are discordant #1 and #2 frequencies equal or skewed?

Quartet Uncertainty (QU)
What proportion of replicates were uninformative? (failed likelihood differential cutoff)

Quartet Fidelity (QF)
When this taxon is sampled, how often does it produce a concordant topology?

Examples:
QC=1 \rightarrow all concordant
QC=0 \rightarrow equivocal conc./disc.
QC<0 \rightarrow disc. > concordant

QD=0 \rightarrow all concordant
QD=0.3 \rightarrow skewed
QD=1 \rightarrow all #1 or #2

QU=0 \rightarrow no uninformative
QU=0.3 \rightarrow 30% uninformative.
QU=1 \rightarrow none informative

Examples:
QF=1 \rightarrow all concordant
QF=0.1 \rightarrow 10\% concordant
QF=0 \rightarrow none concordant
A. Non-vascular plants

B. Ginkgo + cycads

C. Gnetidae + conifers

D. "ANA grade" angiosperms

E. Magnoliids, monocots, and eudicots