RESEARCH ARTICLE

American Journal of Botany

INVITED SPECIAL ARTICLE
For the Special Issue: Using and Navigating the Plant Tree of Life

Allele phasing has minimal impact on phylogenetic reconstruction from targeted nuclear gene sequences in a case study of *Artocarpus*

Heather Rose Kates[1,2,6,*], Matthew G. Johnson[3,4,6,*], Elliot M. Gardner[3,5], Nyree J. C. Zerega[3,5], and Norman J. Wickett[3,5]

Manuscript received 5 September 2017; revision accepted 29 January 2018.

[1] Genetics Institute, University of Florida, P.O. Box 103610, Gainesville, FL 32611, USA

[2] Florida Museum of Natural History, University of Florida, Gainesville, FL 32611, USA

[3] Department of Plant Sciences, Chicago Botanic Garden, 1000 Lake Cook Road, Glencoe, IL 60022, USA

[4] Department of Biological Sciences, Texas Tech University, 2401 Main Street, Lubbock, TX 79414, USA

[5] Plant Biology and Conservation, Northwestern University, 2205 Tech Drive Hogan 2-144, Evanston, IL 60208, USA

[6] Authors for correspondence (e-mails: hkates@ufl.edu; matt.johnson@ttu.edu)

*These authors contributed equally to this work

Citation: Kates, H. R., M. G. Johnson, E. M. Gardner, N. J. C. Zerega, and N. J. Wickett. 2018. Allele phasing has minimal impact on phylogenetic reconstruction from targeted nuclear gene sequences in a case study of *Artocarpus*. *American Journal of Botany* 105(3): 404–416.

doi:10.1002/ajb2.1068

PREMISE OF THE STUDY: Untapped information about allele diversity within populations and individuals (i.e., heterozygosity) could improve phylogenetic resolution and accuracy. Many phylogenetic reconstructions ignore heterozygosity because it is difficult to assemble allele sequences and combine allele data across unlinked loci, and it is unclear how reconstruction methods accommodate variable sequences. We review the common methods of including heterozygosity in phylogenetic studies and present a novel method for assembling allele sequences from target-enriched Illumina sequencing libraries.

METHODS: We performed supermatrix phylogeny reconstruction and species tree estimation of *Artocarpus* based on three methods of accounting for heterozygous sequences: a consensus method based on de novo sequence assembly, the use of ambiguity characters, and a novel method for incorporating read information to phase alleles. We characterize the extent to which highly heterozygous sequences impeded phylogeny reconstruction and determine whether the use of allele sequences improves phylogenetic resolution or decreases topological uncertainty.

KEY RESULTS: We show here that it is possible to infer phased alleles from target-enriched Illumina libraries. We find that highly heterozygous sequences do not contribute disproportionately to poor phylogenetic resolution and that the use of allele sequences for phylogeny reconstruction does not have a clear effect on phylogenetic resolution or topological consistency.

CONCLUSIONS: We provide a framework for inferring phased alleles from target enrichment data and for assessing the contribution of allelic diversity to phylogenetic reconstruction. In our data set, the impact of allele phasing on phylogeny is minimal compared to the impact of using phylogenetic reconstruction methods that account for gene tree incongruence.

KEY WORDS alleles; HybSeq; incomplete lineage sorting; Moraceae; phylogenetics; phylogenomics; target enrichment.

Phylogenetic studies to reconstruct relationships among plant species at all taxonomic scales increasingly rely on hundreds of low-copy nuclear genes. Such biparentally inherited and highly variable nuclear DNA is crucial to reconstruct relationships among closely related species, but these phylogenomic data sets introduce challenges to phylogenetic inference. For example, multiple methods for reconstructing species trees that consider the overall distribution of discordant gene trees have been developed (e.g., Ané et al., 2007;

Liu et al., 2008; Kubatko et al., 2009; Heled and Drummond, 2010); however, a major but mostly overlooked challenge to phylogenetic analysis of phylogenomic data sets is the treatment of heterozygosity within and across intra-individual polymorphic loci. Although intra-individual polymorphisms are rich sources of phylogenetic information, most algorithms for phylogenetic inference treat intra-individual polymorphic sites as ambiguous or missing characters (Potts et al., 2014) that negatively impact phylogenetic resolution.

Standard practice in assembling sequencing reads from heterozygous individuals for phylogenetic studies is to create a single consensus sequence per individual per locus (for exceptions, see Edwards et al., 2008; Weisrock et al., 2012; Kates et al., 2017). In consensus sequences, heterozygous positions are either IUPAC ambiguity-coded or assigned a single nucleotide based on read frequency. In the latter case, the resulting assembled locus is treated as homozygous and is most likely a mixed assembly of two alleles.

A major problem with ambiguity-coded consensus sequences is the handling of ambiguity codes in phylogenetic software. Ambiguity codes are interpreted as missing data by many programs (e.g., BEAST: Drummond and Rambaut, 2007; Mesquite: Maddison and Maddison, 2017; MrBayes: Ronquist and Huelsenbeck, 2003; PAUP*: Swofford, 2002). If ambiguity-coded positions are interpreted as missing data and phylogenetic information in a heterozygous position of an alignment is lost, phylogenetic resolution may be lower among highly heterozygous sequences than among sequences with less genetic variation. Exceptions to this treatment of ambiguity-coded positions as missing data are implemented RAxML (Stamatakis, 2006), ape (Paradis et al., 2004), and SVDquartets (Chifman and Kubatko, 2014). These programs all have options to treat ambiguity-coded positions as informative, whether as polymorphisms (multiple states present; e.g., RAxML) or as true ambiguities (one of multiple, possible states present; e.g., ape). When ambiguity codes are treated as polymorphisms in likelihood models, the probability of substituting an "A" by "Y" equals that of "A" by "C" and/or "T", which increases computation time and topological uncertainty (Felsenstein, 2003). The issues associated with ambiguity-coded bases are particularly problematic for data sets in which individuals are highly heterozygous but allelic diversity at the population, species, or genus level is low (Potts et al., 2014).

The most biologically accurate way to include heterozygous genes in phylogenetic analysis is to reconstruct phylogenetic trees from allele sequences rather than from ambiguity-coded consensus sequences or chimeric consensus sequences. This process, known as "phasing," joins variants across sites: for example, the "G" in a G/T single-nucleotide polymorphism (SNP) at one site may be associated with the "C" in a C/A SNP at another site. Two main methods of generating phased allele sequences for a single locus are available: amplicon sequencing, which can avoid the need for phasing altogether, and statistical phase inference. In amplicon sequencing (e.g., Uribe-Convers et al., 2016; Kates et al., 2017; Rothfels et al., 2017), targeted genes are amplified in separate PCR reactions and pooled for sequencing. If sequencing read length is longer than the PCR product, reads will have identical start and end sites and may be separated into two or more alleles without phasing. However, amplicon sequencing requires a large amount of wet-lab preparation to produce a phylogenomic data set. Furthermore, to assemble two or more alleles that do not require phasing, PCR primers must target relatively short loci unless long read sequencing is used (e.g., Rothfels et al., 2017), and custom read assembly is required.

There are two main methods for inferring phase statistically: population-based phasing and read-backed phasing. The former requires a reference population of known variants, where the phase between alleles is already known, and is used almost exclusively for well-characterized model systems (e.g., Browning and Browning, 2007). In read-backed phasing, reads are aligned to a reference sequence (a genome reference or the consensus sequence produced by de novo assembly), and variant sites (i.e., SNPs) are detected. If variants are connected by read data, they can be phased into short-range haplotypes (Fig. 1). This method is limited by read length and depth, especially in targeted sequencing projects. Intergenic regions with zero read coverage preclude phasing all variants into long-range haplotype blocks, and coverage within long introns may create multiple phaseable regions within each locus (Fig. 1).

Even if a data set of allele sequences is assembled, phylogeny reconstruction from two sequences per individual is also not straightforward. Common methods for phylogeny reconstruction from large numbers of nuclear genes include concatenation (i.e., supermatrix) and species tree estimation (e.g., ASTRAL: Mirarab and Warnow, 2015; *BEAST: Heled and Drummond, 2010). There is no clear way to concatenate allele sequences across loci without long-range haplotype phase information, and multiple studies have demonstrated that when a single allele for each heterozygous locus is selected and included in the concatenated data matrix, the selection of alleles can influence the results (see Edwards et al., 2008; Weisrock et al., 2012). The influence of allele selection on the results occurs, in part, because when incomplete lineage sorting occurs, heterozygous alleles from one species may coalesce deeper in a particular gene phylogeny than that species' divergence with its sister species (Weisrock et al., 2012).

Species tree methods may be used to address issues with discordant gene histories due to deep coalescence (Knowles, 2009; Liu et al., 2009a, 2015; Mirarab and Warnow, 2015). Because these methods allow for mapping of multiple "individuals" in gene trees or a sequence data matrix to a single "species" in the species tree, they include a practical way to use allele information: multiple alleles in gene trees are mapped to a single individual in the species tree.

Standard methods to assemble and analyze sequence data do not enable researchers to easily reconstruct the evolutionary history of alleles, so we do not know the extent to which phylogeny reconstruction from allele sequences could more effectively elucidate phylogenetic relationships that remain unknown. Here, we present a method for assembling an allele data set from Illumina (San Diego, CA, USA) sequencing reads and describe how to integrate this method into a recently developed tool for locus assembly from target enrichment data, HybPiper (Johnson et al., 2016). We demonstrate how these data can be used for phylogeny reconstruction and compare phylogenetic hypotheses of *Artocarpus* J.R.Forst. & G.Forst (Moraceae) (breadfruit, jackfruit, and relatives) across allele and non-allele data sets using multiple methods of phylogenetic inference to determine whether the use of allele data influences phylogenetic results.

METHODS

Data sets

The *Artocarpus* data set includes target-enriched sequences of 23 ingroup taxa and one outgroup taxon from the sister tribe Moreae (*Steblus glaber* Corner) (Table 1), originally described in detail by

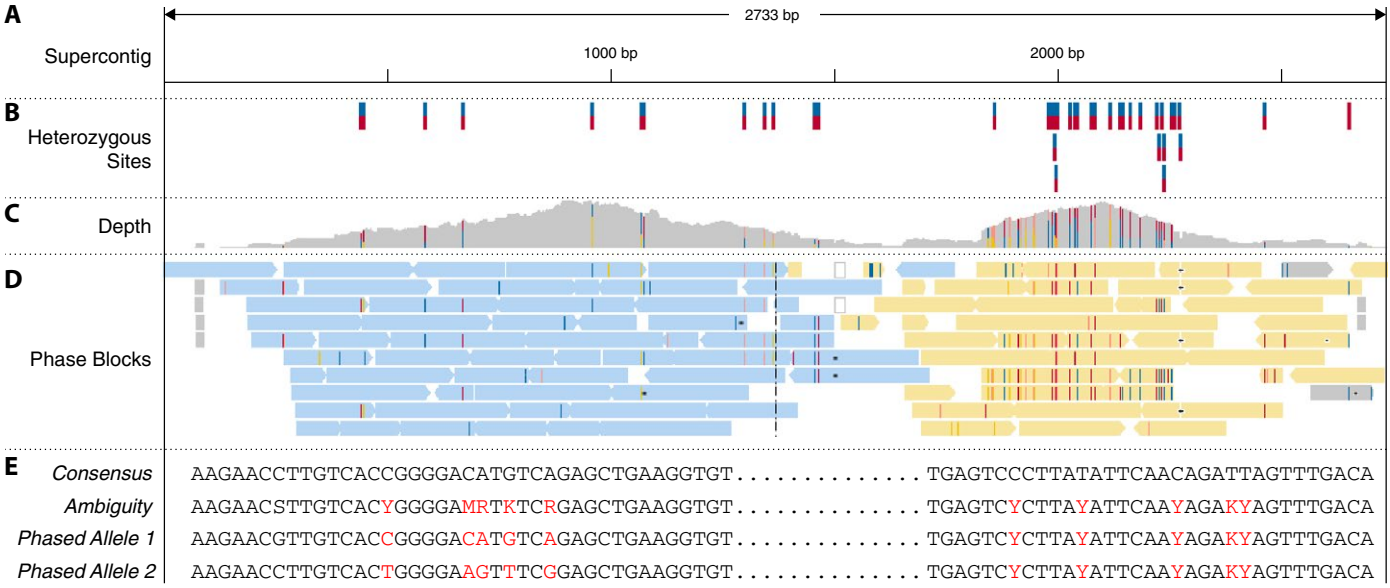

FIGURE 1. Read-backed phasing of targeted loci from high-throughput sequences. (A) The HybPiper supercontig (concatenated exons with flanking introns) for one locus (gene250.pl) for *Artocarpus lowii* (MWL2). (B) Heterozygous sites identified by GATK are marked with vertical bars; the proportion of blue to red reflects the proportion of reads identified with each allele. (C) Histogram of sequencing depth (varies from 1× to 45×) following the removal of duplicate reads using Picard. (D) Read-backed phasing using WhatsHap separated reads into two phase blocks, separated by a gap of low depth. Variant sites in the left block (blue) can be phased with other sites within the block; similarly, variants in the right block (yellow) can be phased with respect to other sites in the yellow region. However, variants in the blue block cannot be phased with respect to variants in the yellow block, because they are not connected by sufficient read data. (E) Partial DNA sequence generated by haplonerate.py for the three data sets for the same locus and individual. Phased alleles (red text) were retained only in the largest phase block (left).

TABLE 1. Sample information, locus recovery, and statistics for allelic variation for 23 species of *Artocarpus* and one outgroup. The percentage of longest phase block refers to genes that have more than one phase block for that individual. Voucher information for all individuals except MV2 (vouchered in Gardner et al., 2016) can be found in Johnson et al., 2016.

Species	Subgenus	Sample ID	No. loci	Loci with two alleles	Loci with >1 phase block	Percentage longest phase block	Genes with deep coalescence of alleles	Mean Sqrt pairwise allelic distance
Artocarpus anisophyllus	*Artocarpus*	NZ606	111	90	22	75.2%	10	0.0676
Artocarpus brevipedunculatus	*Artocarpus*	NZ814	111	89	22	76.0%	4	0.0681
Artocarpus camansi	*Artocarpus*	MV2	111	28	1	91.2%	1	0.0548
Artocarpus elasticus	*Artocarpus*	EG87	111	77	14	83.0%	4	0.0625
Artocarpus excelsus	*Artocarpus*	NZ780	111	81	22	79.2%	4	0.0600
Artocarpus kemando	*Artocarpus*	NZ612	111	85	18	78.6%	3	0.0625
Artocarpus lanceifolius	*Artocarpus*	NZ739	111	85	15	74.2%	10	0.0711
Artocarpus lowii	*Artocarpus*	MWL2	111	62	17	76.2%	2	0.0617
Artocarpus odoratissimus	*Artocarpus*	NZ866	111	89	21	72.6%	3	0.0744
Artocarpus rigidus	*Artocarpus*	NZ728	111	79	25	77.9%	3	0.0721
Artocarpus sericicarpus	*Artocarpus*	NZ771	111	72	20	78.0%	2	0.0644
Artocarpus tamaran	*Artocarpus*	EG92	111	91	23	75.9%	4	0.0690
Artocarpus teysmannii	*Artocarpus*	NZ946	111	37	6	74.3%	1	0.0644
Artocarpus sepicanus	*[Artocarpus]*	GW1701	109	77	27	77.7%	2	0.0712
Artocarpus heterophyllus	*Cauliflori*	EG98	111	48	17	71.9%	3	0.0817
Artocarpus integer	*Cauliflori*	NZ918	111	91	29	74.1%	1	0.0737
Artocarpus limpato	*Prainea*	NZ609	109	81	21	76.2%	0	0.0656
Artocarpus fretessii	*Pseudojaca*	NZ929	110	94	22	72.5%	12	0.0774
Artocarpus lacucha	*Pseudojaca*	NZ420	111	77	24	78.1%	4	0.0706
Artocarpus nitidus ssp. Lingnanensis	*Pseudojaca*	NZ911	111	29	5	83.5%	5	0.0805
Artocarpus peltatus	*Pseudojaca*	NZ694	111	90	24	76.3%	7	0.0815
Artocarpus primackiana	*Pseudojaca*	NZ687	110	93	27	74.5%	6	0.0734
Artocarpus thailandicus	*Pseudojaca*	NZ402	110	69	16	71.7%	23	0.1320
Streblus glaber	NA	EG78	111	71	26	78.6%	0	0.0623

Note: Based on phylogenetic evidence, *Artocarpus sepcianus* likely does not belong in *Artocarpus* subg. *Artocarpus* (Johnson et al., 2016, this paper)

Johnson et al. (2016). The target enrichment probes (MycroArray, Ann Arbor, MI, USA) were originally designed for 333 single-copy loci; however, many of these genes have one or more paralogs, resulting from the whole genome duplication in *Artocarpus* (Gardner et al., 2016). To avoid potential problems with assembling paralogs as alleles, we built new assemblies for the present analyses as follows. Target-enriched reads from *A. camansi* Blanco (the same individual used for whole-genome sequencing in the original marker development (Gardner et al., 2016)) were assembled de novo using SPAdes (Bankevich et al., 2012), and genes were predicted using Augustus (Keller et al., 2011), with *Arabidopsis* Hehyn. as the reference. Those genes were annotated using a BLASTn search seeded with the HybPiper target file of 458 genes from Johnson et al. (2016). We previously demonstrated that a majority of the "phylogenetic" genes in the target set (low copy genes, in contrast with high-copy MADS-box and volatile genes in the target set) have multiple paralogs in *Artocarpus* due to a whole-genome duplication in the common ancestor of the genus (Gardner et al., 2016). We constructed a new target file containing paralogs of each gene (726 total loci) to use for HybPiper. Because additional paralog warnings within this set suggested further clade-specific gene duplications within *Artocarpus*, for this study, we chose to focus only on genes that had only one copy in *Artocarpus*: 151 genes that never triggered a HybPiper paralog warning for any sample. Paralog warnings are triggered when multiple contiguous sequences are assembled within HybPiper that each represent more than 85% of the targeted sequence length. From among these genes, we selected 111 genes with the outgroup present for the analyses presented here.

We assembled four data sets from the sequencing reads: (1) consensus sequences with heterozygous bases called as the nucleotide with highest read-frequency ("*consensus*"), (2) consensus sequences with heterozygous bases ambiguity-coded ("*ambiguity*"), (3) allele sequences with read-backed phasing ("*alleles*"), (4) unphased allele sequences ("*unphased alleles*") used in SVDQuartets (Chifman and Kubatko, 2014) only. All other phylogenetic methods described below were used only for the *consensus*, *ambiguity*, and *alleles* data sets.

We used HybPiper to assemble the *consensus* data set, which is the default method of sequencing read assembly implemented in HybPiper and is described in detail in Johnson et al. (2016). Briefly, HybPiper sorts reads by target gene and then performs de novo assembly of contigs for each gene using SPAdes (Bankevich et al., 2012). If more than one long contig is assembled by SPAdes, HybPiper chooses among multiple full-length contigs by using a sequencing depth cutoff to choose the best full-length contig. If the sequencing depth is similar among all full-length contigs, HybPiper chooses one based on the percent identity with the reference sequence. This method results in a single contig, even if multiple long sequences represent alleles (Johnson et al., 2016). We extracted exon sequence and flanking regions for each gene, which were concatenated into a "supercontig" using the "intronerate.py" script in HybPiper.

We assembled the *ambiguity* data set using the following steps for each sample (the total 726 sequenced genes were used for assembly purposes): (1) We used BWA-MEM (Li, 2013) and Picard (https://broadinstitute.github.io/picard/) to generate a duplicate-free BAM alignment for each sample separately, using the "supercontig" sequences generated by HybPiper as a reference, (2) we used GATK (McKenna et al., 2010) to identify and call variants, retain SNPs (using a hard filter that ensures SNPs are identified using standard quality and depth thresholds) and generate sequences with IUPAC ambiguity codes, and (3) for each gene, we generated separate exon and intron files using annotations created by HybPiper (Johnson et al., 2016). (4) We used Macse (Ranwez et al., 2011) to generate in-frame alignments for exons and used MAFFT (Katoh and Standley, 2013) to align intron sequences separately. (5) We trimmed exon and intron alignments with TrimAl (Capella-Gutiérrez et al., 2009) to remove sites that did not appear in at least 15 samples and combined the alignments retaining positional information for gene partitions (Appendix S1A, see Supplemental Data with this article).

To assemble the *alleles* data set, we used WhatsHap (Patterson et al., 2015), a Python-based program for read-backed phasing designed for long-reads but appropriate for 300-bp paired-end Illumina data. For each sample, we used the BAM alignments generated for the *ambiguity* data set to generate a phased VCF file and a GTF file containing the locations of phase blocks within each gene using default WhatsHap settings. We generated two phased sequences for every gene using "bcftools consensus", but we only retained phased sequence in the longest phase block for each individual at each gene. The remaining variant sites outside the longest phase block were replaced with ambiguity characters (Fig. 1), and only one sequence was retained if it was completely homozygous for that individual. This method allowed us to retain phased heterozygous sites only where backed by read data (i.e., not spanning long introns), while retaining the full-length sequence that may contain informative sites across species. Our script for processing phased alleles, "haplonerate.py" is freely available at www.github.com/mossmatters/phyloscripts

The final product from the *alleles* assembly method is one or two sequences per individual per gene. If an individual is heterozygous at a given locus, assembly of allele sequences results in intra-individual alleles: two alleles (for a gene) from the same individual. If intra-individual alleles are not sister in a gene tree, we refer to this as "deep coalescence of alleles". Finally, we assembled a second version of the alleles data set ("*unphased*"), which always contains two sequences per individual with no ambiguity codes. We used this assembly to test whether phasing alleles was appropriate for SVDQuartets (Chifman and Kubatko, 2014; described next in *Phylogenetic analyses*), a method where each site is treated independently.

Phylogenetic analyses

For the *ambiguity*, *consensus*, and *alleles* data sets, we estimated gene trees with RAxML v 8.2.3 (Stamatakis, 2014) using the GTR+Gamma model of evolution and 1000 bootstrap replicates. We specified separate partitions for (1) first and second codon position, (2) third codon position, and (3) intron for each gene.

For the *ambiguity* and *consensus* data sets, we performed concatenation + maximum likelihood analysis (CAML) on concatenated alignments of 111 genes using RAxML with the GTR+Gamma model of evolution and 1000 replicates of multilocus bootstrapping. We partitioned the concatenated alignments by gene, codon position, and intron (330 partitions per alignment). There is no known method for treating the *alleles* data set in a CAML analysis, because alleles would need to be associated across loci into long-range haplotypes.

We estimated species trees using four popular summary methods that are consistent under the multi-species coalescent:

ASTRAL II v 4.7.4 (Mirarab and Warnow, 2015), MP-EST v 1.5 (Liu et al., 2010), STAR (Liu et al., 2009b), and SVDquartets (Chifman and Kubatko, 2014). We used ML gene trees estimated by RAxML (described above) as input for ASTRAL-II, MP-EST, and STAR. For each data set, we ran ASTRAL-II multi-locus bootstrapping (option -g) with 1000 bootstrap replicates (option -r). For the *alleles* data set, we used the mapping option (-a) to map alleles in the gene trees to individuals. Because branches with very low support in gene trees may cause erroneous topologies in the species tree, we also estimated the ASTRAL species trees using ML gene trees with branches with less than 33% bootstrap support (BS) collapsed using TreeCollapserCL4 (http://emmahodcroft.com/TreeCollapseCL.html). We could not collapse branches with low support for the other species tree methods as they require bifurcating trees.

We estimated species trees and performed bootstrap validation with MP-EST v 1.5 and STAR on STRAW (http://bioinformatics.publichealth.uga.edu/SpeciesTreeAnalysis/index.php). Before MP-EST and STAR analysis, we rooted ML gene trees and bootstrap trees by outgroup using the program pxrr in phyx (Brown et al., 2017) and removed branch lengths and support values from the trees. We added BS values from the bootstrap consensus trees to the species trees manually. For the *alleles* data set, we used a species-allele table to map alleles in the gene trees to individuals.

We estimated SVDquartets trees from concatenated gene sequence alignments using SVDquartets analysis in PAUP* version 4.0a158 (Swofford, 2002) with exhaustive quartet sampling, 1000 bootstrap replicates, and the multispecies coalescent tree model. For the *ambiguity* data set, we ran one analysis with ambiguity-codes interpreted as distributed and one analysis with ambiguity-codes considered as missing data. For the *alleles* data sets, species-membership partitioning was used to assign allele sequences to individuals. We also performed SVDquartets analysis on unphased allele sequences to determine whether this method affects the SVDquartets tree from allelic data. SVDquartets treats each alignment position as an unlinked locus, so phasing of SNPs across alleles may not be necessary for this type of species tree analysis. We manually added bootstrap support values from the bootstrap consensus trees generated by SVDquartets to the SVDquartets species trees.

To focus on issues that may arise in phylogeny reconstruction in data sets with a high level of deep coalescence of alleles, we performed additional species tree estimation in ASTRAL and CAML analysis on a subset of seven genes that had the highest proportion of nonsister intra-individual alleles (20% or higher) in the ML gene trees. The identification of these genes is described below, and we performed the ASTRAL and CAML analyses exactly as described above for the full data sets (ASTRAL using ML gene trees with branches with less than 33% BS support collapsed). We estimated "percent resolution" of the resulting trees for each data set by collapsing branches in the trees that had <50% BS support using TreeCollapser4 ((http://emmahodcroft.com/TreeCollapseCL.html) and counting the number of bipartitions in the resulting trees using the function bitsplits() in the R package ape (Paradis et al., 2004). We then calculated percentage resolution as the number of bipartitions in these trees divided by the number of possible bipartitions (two less than the number of tips). For the *alleles* data set, we removed all nodes where the descendants were two alleles from the same individual, to avoid overestimating support on these gene trees.

Assessing gene tree resolution, topological incongruence, and allele coalescence

We used Phyparts (Smith et al., 2015; https://bitbucket.org/blackrim/phyparts) to assess gene tree discordance and gene tree-species tree discordance. Phyparts conducts bipartition analysis across a set of trees while allowing for missing data. Phyparts bins gene tree nodes into four categories relative to a reference tree, including a category for those that inform (conflict or support) a clade but have less than 50% bootstrap support. For each clade in the reference tree, gene trees in this category were not included in our comparisons of the number of gene trees concordant with the species tree between the consensus and ambiguity data sets. For each of the three data assemblies, we ran phyparts using the ASTRAL-II species tree (estimated from that data set) as the mapping tree to assess discordance among the gene trees. Although we include a phyparts analysis of the gene trees estimated from the alleles data set, the concordance values are not directly comparable to the other data sets for two reasons: (1) To compare gene trees in which all allele sequences are tips to a species tree, the species tree must also have one tip per allele sequence. Because intra-individual alleles are arbitrarily named across gene trees, the only appropriate way to estimate a species tree for from multiple alleles gene trees is to map alleles to individuals. The *alleles* ASTRAL-II species tree used for the mapping tree is therefore not a real estimation of relationships. (2) Any nonconcordance between a gene tree and the ASTRAL-II species tree that involves the position of an allele that is not sister to its second intra-individual may not be reflective of arbitrary intra-individual allele naming. Although any interpretation of phyparts results for the *alleles* data set is severely limited, we include the general patterns observed because the incidences of nonsisterhood of intra-individual alleles was rare in most gene trees after we collapsed branches with low support (described in results).

Before running phyparts, we rooted gene trees and the mapping tree by outgroup using the pxrr program in phyx (Brown et al., 2017) and removed branch lengths from the mapping trees. We ran phyparts with the -b option set to 33 so that branches with less than 33% BS support in the gene trees would not be considered. Results from phyparts were visualized using phypartspiecharts.py to summarize gene tree conflict on our phylogenies using pie charts. We used a second script (minorityreport.py) to generate a "minority bipartition" report from the phyparts output that shows the number of gene trees supporting each alternative bipartitions. Both scripts are freely available under an MIT license available at the website www.github.com/mossmatters/phyloscripts.

To summarize the support for hypotheses of *Artocarpus* relationships across data sets and analyses, we compared the topologies of the 17 species trees and two CAML trees by eye after rooting by outgroup using the pxrr program in phyx (Brown et al., 2017). We identified an across-analysis incongruence and manually identified all various arrangements in areas that exhibited incongruence.

To identify deep coalescence of intra-individual alleles within gene trees, we used the Python package ETE3 (Huerta-Cepas et al., 2016); etetoolkit.org) to calculate how often intra-individual alleles were resolved as sister lineages in ML gene trees estimated from the *alleles* data set. When the two intra-individual alleles were monophyletic on a gene tree, we recorded the gene tree bootstrap support. When the alleles were not monophyletic, we recorded the support from the most highly supported node that prevented the monophyly of alleles. We summarized the support for monophyletic

intra-individual alleles values across samples and loci in a heatmap using the Python package seaborn (Waskom et al., 2014). A python notebook describing our procedure is available at the website www.github.com/mossmatters/phyloscripts.

To characterize the level of sequence heterozygosity for each individual, we calculated the square root pairwise distance (sqrt PWD) between intra-individual allele sequences in each of 111 gene alignments from the alleles data set using the function dist.alignment() in the R package seqinR (Charif and Lobry, 2007). For homozygotes, we replicated sequence names and sequences in the alignments to generate sqrt PWDs of zero. To see whether there was an association between the prevalence of deep coalescence of intra-individual alleles for a gene or highly heterozygous genes and gene tree resolution, we plotted the number of nonsister intra-individual alleles in a gene tree and the average sqrt PWD for a gene against gene-tree percent resolution. We estimated "percent resolution" of gene trees as described above in *Phylogenetic analyses.*

RESULTS

Assemblies and alignments

We reduced our gene sampling to maximize taxon occupancy; most species had all 111 genes present, and the fewest genes any taxon had was 109 (Table 1). The gene alignments ranged from 1067 bp to 5350 bp with a median length of 2303 bp. The final concatenated assembly contained 361,738 bp and had only 71 missing gene/individual combinations.

Allele coalescence, sequence heterozygosity, and gene tree resolution

Among the 24 individuals, the number of gene trees in which intra-individual alleles were not sister (i.e., deep coalescence of alleles) ranged from zero to 23 of 111 gene trees, and the average was 4.75 of 111 gene trees (Fig. 2). Across 111 gene trees, the percentage of non-sister intra-individual alleles (heterozygous allele pairs not sister in the tree/total heterozygous allele pairs in the tree) ranged from 0% to 33% and the average was 6.7%. We did not observe an association between the proportion of monophyletic intra-individual alleles (in the *alleles* data set) in a gene tree and the gene tree's percent resolution (in any data set, Fig. 3).

Among the 24 individuals, the sqrt PWD ranged from 0.055 to 0.132 and the average was 0.080 (Appendix S1A). Across 111 genes, the average (for all individuals) sqrt PWD ranged from 0.037 to 0.144, and the average was 0.070. We did not observe an association between individuals that had high heterozygosity and membership in high-conflict/poorly resolved areas of the *Artocarpus* phylogeny (Appendix S1B), nor did we observe an association between a high number of heterozygous sequences in a gene tree and the gene tree's percent resolution (Appendix S1C).

The mean percent resolution across gene trees was 77% for the *ambiguity* data set, 78% for the *consensus* data set, and 86% for the alleles data set (Fig. 3). Paired *t*-tests of gene tree resolution among the three assembly methods were not significant. Gene tree percent resolution for the *alleles* data set is somewhat inflated because of the contribution of strong support for intra-individual allele sisterhood (as many as 12 bipartitions per gene tree) in the alleles gene trees.

Levels of gene tree/species tree discordance across methods of data assembly

The number of gene trees concordant with the species tree was similar for the *consensus* and *ambiguity* data sets for most clades (Fig. 4A, B). Of the 11 nodes that had a high level of gene tree discordance with the species tree (>50% of gene trees discordant with species tree), the number of discordance gene trees differed between the two data sets by more than five for only two nodes (1 and 7). For these two nodes and six of the eight remaining "high-conflict" nodes, the *ambiguity* data set had a higher number of gene trees concordant with the species tree than did the *consensus* data set.

Comparison of gene tree discordance in the *alleles* data set to the *ambiguity* and *consensus* data sets is limited as described in methods. We did not make comparisons of *alleles* gene tree concordance to *ambiguity* and *consensus* species trees at high conflict nodes 1–7 because these subtend clades that include individuals with intra-individual alleles that were nonsister in 10 or more gene trees: NZ402 and NZ929. (The other two individuals with nonsister alleles in 10 or more gene trees occur in parts of the

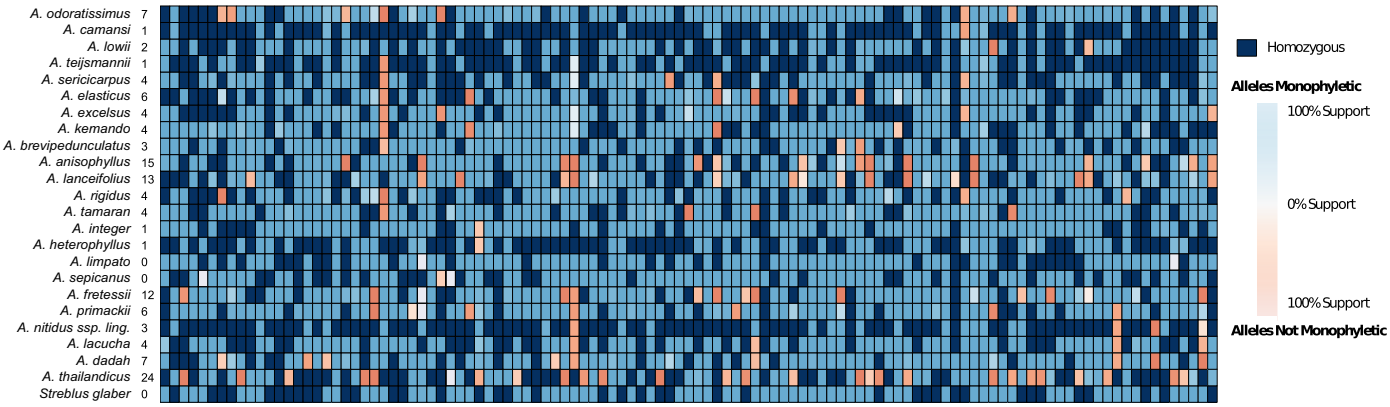

FIGURE 2. Frequency of deep coalescence of intra-individual alleles. Each gene is a column; each row is an individual. The color ranges from light blue (100% support for monophyly of intra-individual alleles) to white (0% support) to red (100% support for non-monophyly of intra-individual alleles). Homozygous loci are shown in dark blue.

tree without high gene tree/species tree discordance.) At high-conflict nodes 16–21 (subtending clades that include no individuals with nonsister alleles in more than four gene trees) we observed no association between the number of gene trees concordant with the species tree and the use of allele sequences for gene tree reconstruction (Fig. 4A and B). Numbers of gene trees concordant with the species tree at these nodes are either lower, in the middle, or higher compared with the *consensus* and *ambiguity* data sets. The total number of alternative bipartitions found in gene trees was similar for each high-conflict node across all three data sets; only one node (node 6) had a difference in the number of alternative bipartitions greater than three between the *ambiguity* and *consensus* data sets.

Topological incongruence among data sets and phylogenetic inference

We assessed incongruence among 19 topologies resolved by all combinations of data sets and phylogenetic analyses: six species trees for the *alleles* assembly data set (Appendix S2.1–S2.6), five species trees and one CAML tree for the *consensus* data set (Appendix S2.7-S2.12), and six species trees and one CAML tree for the *ambiguity* data set. (Appendix S2.13–S2.19). We identified four major areas in the *Artocarpus* phylogeny where species' placement varied among the 19 trees (the two most salient detailed in Fig. 5, and all four detailed in Appendix S3). Hypotheses for *Artocarpus* evolution are discussed extensively elsewhere (Williams et al., 2017). Here, we simplify our descriptions of alternative hypotheses to focus on comparing results across data sets.

Briefly, the four areas of incongruence are (A) the position of subgenus *Prainea* (King) Renner, here represented by *A. limpato* Miq. (NZ609) (Fig. 5). *Prainea* occurs successively or as a two-member clade with *A. sepicanus* Diels as first-branching taxa to the rest of the genus or to a major clade, or as first-branching species in different major clades. (B) Series *Rugosi* Jarrett of subgenus *Artocarpus* sect. *Artocarpus* (Fig. 5) (also including *A. teijsmannii* Miq.), involving comparatively shallow nodes. Three species (*A. tamaran* Becc. [EG92], *A. sericarpus* F.M.Jarret [NZ711], and *A. elasticus* Reinw. ex Blume [EG87]) occur in all possible positions within the clade. (C) Resolution within subgenus *Pseudojaca* Tréc. (Appendix S3). Four of the six subgenus *Pseudojaca* species [*A. dadah* Miq. NZ694), *A. thailandicus* C.C.Berg (NZ402), *A. nitidus* subsp. *lingnanensis* (Merr.) F.M.Jarrett (NZ911), *A. lacucha* Buch.-Ham. (NZ420)] occur in every possible position within the subgenus; (D) The positions of the allied species *A. excelsus* F.M.Jarrett (NZ780) an *A. lowii* King (MWL2), which either form a clade sister to or a grade before Area B (Appendix S3).

The crown age of the ingroup has been estimated at ca. 40 Myr, although several subclades contain species splits as young as <5 Myr ago (Ma) (Williams et al., 2017). Branches involved in incongruent relationships in Area A (crown age ca. 40 Myr) are

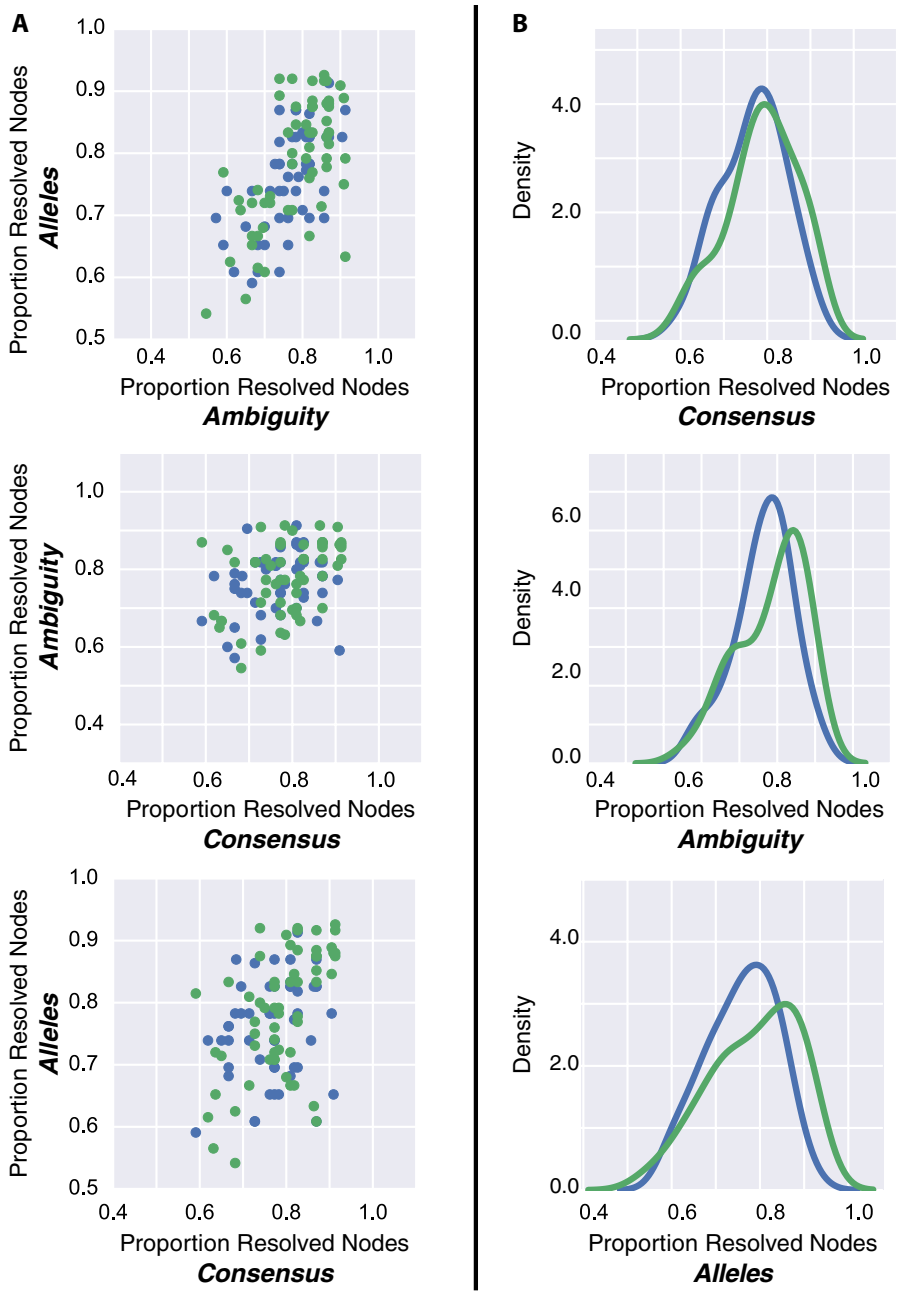

FIGURE 3. Relationship between deep coalescence and gene tree resolution. Resolution is measured by the percentage of nodes with >50% support. Gene trees with evidence of deep coalescence (non-monophyly of interspecific alleles) are in green; genes without deep coalescence are in blue. (A) Pairwise comparisons of gene tree resolutions between methods. (B) Univariate kernel density distribution of gene tree resolution for each of the three assembly methods.

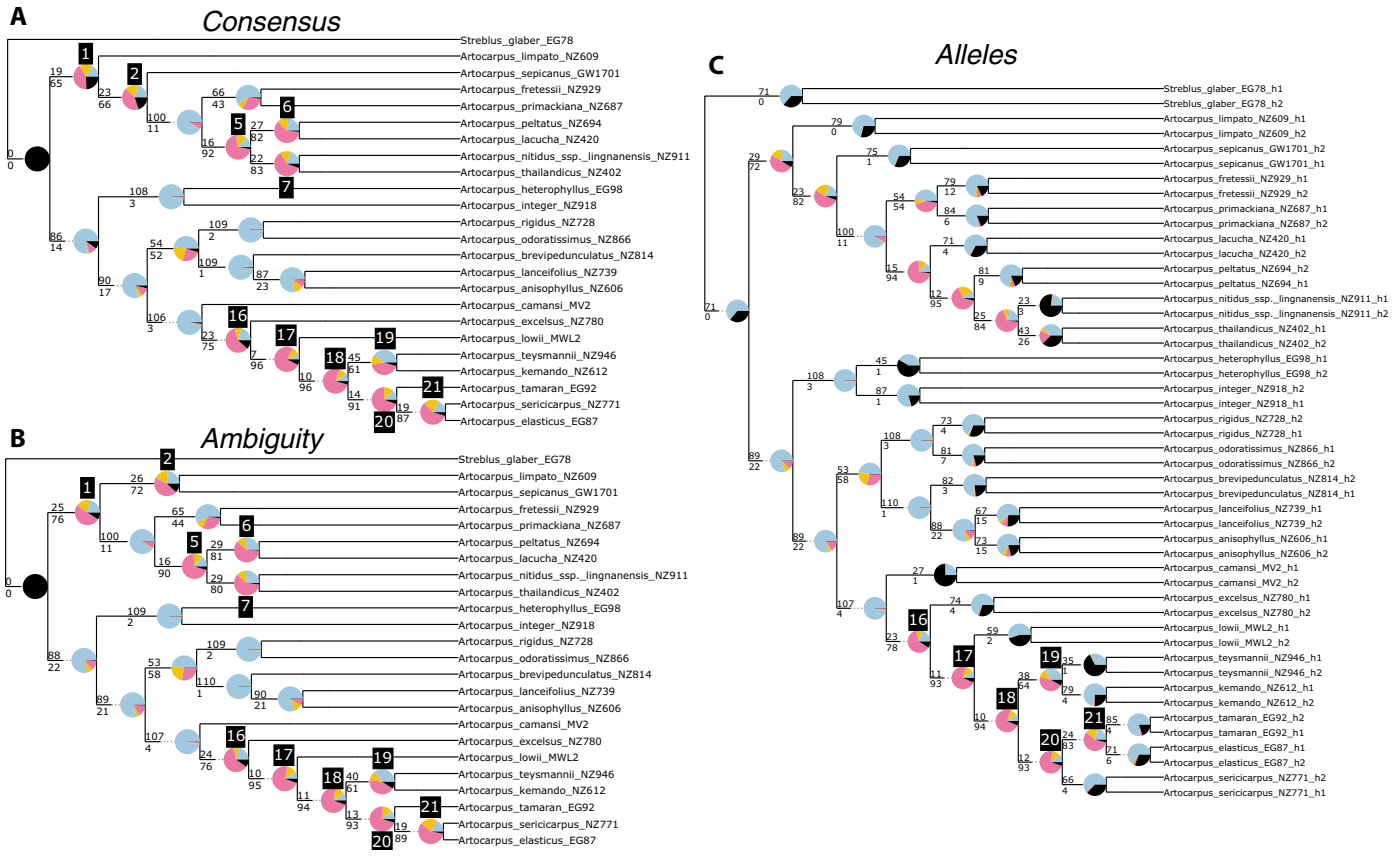

FIGURE 4. ASTRAL-II trees for *Artocarpus*. (A) *Consensus* data set; (B) *ambiguity* data set; (C) *alleles* data set with summary of conflicting and concordant gene trees produced with Phyparts. For each branch, the top number indicates the number of gene trees concordant with the species tree at that node, and the bottom number indicates the number of gene trees in conflict with that clade in the species tree. The pie charts at each node present the proportion of gene trees that support that clade (blue), the proportion that support the main alternative for that clade (yellow), the proportion that support the remaining alternatives (pink), and the proportion that inform (conflict or support) this clade that have less than 50% bootstrap support (black). Numbers in black boxes indicate nodes of interest discussed further in the text. Alternative topologies and the number of gene trees that support them can be found in the data repository.

the deepest in the tree, and branches affecting the resolution of Area D (crown age ca. 15 Myr) are the next deepest in the tree. Branches in Areas B (crown age ca. 15 Myr, but here undoubtedly younger than D due to topological differences with Williams et al. [2017]) and C (crown age ca. 13 Myr) are at comparably shallow depths in the tree and subtend terminal clades and/or one node deeper depending on the specific topology resolved. Of the four areas of incongruence, only Area C contained samples with high rates (>10% of genes) of nonsister alleles (*A. thailandicus* and *A. fretessii*) (Fig. 2). By contrast, the other two samples with similarly high levels of intraspecific deep coalescence (the sister species *A. lanceifolius* and *A. anisophyllus*, members of a clade with a crown age of ca. 13 Myr) were not associated with an area of phylogenetic incongruence.

We found slightly more variation in topologies resolved between data sets (within each program) than between programs (within data sets) (Table 2). In Area A, six arrangements occurred in our trees (Fig. 5). Area A was resolved more congruently between the *alleles* and *consensus* data sets than between either of these and the *ambiguity* data set, and these two data sets resolved trees with generally higher BS support in this area of the tree. In 16 of 19 trees, *Prainea* was not sister to all other *Artocarpus* species but was instead nested inside the genus, agreeing with the treatment of *Prainea* by Zerega et al. (2010) as a subgenus of *Artocarpus*.

In Area B, four arrangements of ser. *Rugosi* occurred in our species/CAML trees, but only one arrangement was resolved with high BS support (>70%) in more than one tree (Fig. 5). The arrangement resolved in the majority of trees was resolved with similar frequency by all of the data sets. The *consensus* data set had the most between-analyses congruence in area C, and area C was resolved more congruently between the *alleles* and *ambiguity* data sets than between either of these and the *consensus* data set. By contrast with area B, this clade was either mostly (Zerega et al., 2010) or completely (Berg et al., 2006) sampled for species, depending on the treatment.

In Area C, eight arrangements for subg. *Pseudojaca* occurred in our trees, but only one arrangement was resolved with high BS support (>70%) in more than one tree (Appendix S3). This most common arrangement was resolved consistently in species trees for both the *ambiguity* and *consensus* data sets, but there was complete incongruence among analyses for the *alleles* data sets; each analysis resolved a different topology. Indeed, four of the eight arrangements appeared only in the *alleles* analyses. Area B was resolved more congruently between the *consensus* and *ambiguity* data sets than between either of these and the *alleles* data

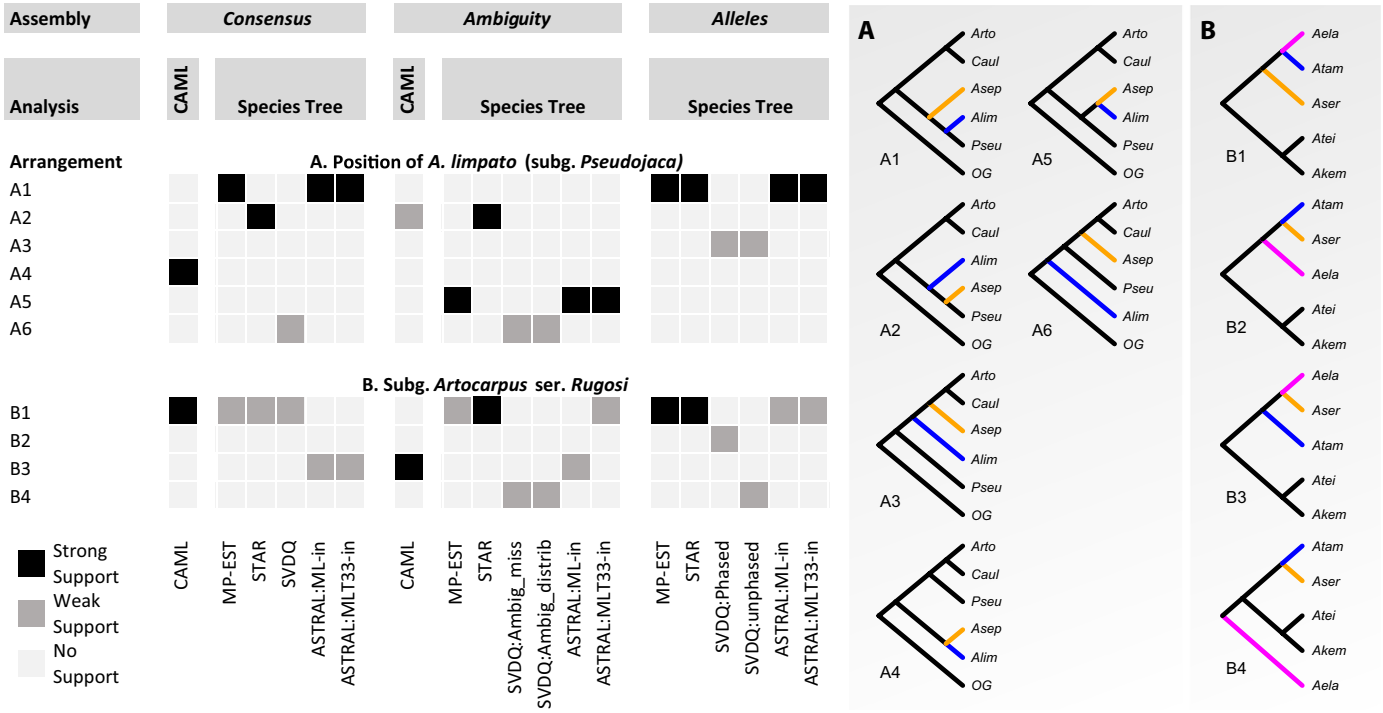

FIGURE 5. Summary of species tree conflict between analyses for all data sets for two major regions of incongruence. Left: Table showing incongruence between analyses for all three data sets for Areas A and B. Rows represent topologies, drawn as cartoon trees (or subtrees) at right. Columns represent analyses (labeled at bottom) within data sets (labeled at top). Strong support (black) >70%, weak support <70%. No support indicates that the topology in question was not recovered. Right: Cartoon trees representing alternative topologies: (A) position of *A. limpato* (subg. *Prainea*), with other subgenera collapsed; (B) subg. *Artocarpus* ser. *Rugosi*. Rearrangements are shown by colored terminal branches within each area.

set. This clade was sparsely sampled in our data set, with only 6 of 24 species represented.

In Area D, three arrangements for *A. excelsus* and *A. lowii* occurred, and two arrangements had high (>70%) BS support in six of 19 trees (the third arrangement only occurred in one tree and with low support) (Appendix S3). The *consensus* data set had the highest among-analyses congruence in this area. Area D was resolved more congruently between the *ambiguity* and *allele* data sets than between either of these and the *consensus* data set because of the congruence between the SVDQ trees for these data sets. This was the only high-conflict area in the tree for which the CAML analyses resolved congruent topologies for the *consensus* and *ambiguity* data sets.

Comparison of support for relationships reconstructed using different data sets

For the four areas of incongruence, we compared the BS support across trees for the first-branch if the arrangement varied by first-branching species or for the base of the clade if the arrangement varied by clade membership (Fig. 5; Appendix S3). (We did not compare BS support for areas without high conflict in the tree because these areas had high support in all trees.)

For the *consensus* data set, the CAML tree had high BS support for the arrangements it resolved in all four high-conflict areas. Each ASTRAL tree had high support for the arrangements it resolved in

TABLE 2. Topological variation among three data sets (A) and within each data set (B) for four areas of species tree incongruence (see Appendix S3). Percentages are number of pairwise differences across trees/possible number of pairwise differences across trees; higher numbers indicate greater topological variation.

A. Percentage variation among data sets (within analyses)					
Area of incongruence	**ASTRAL**	**MPEST**	**STAR**	**SVDQ**	**CAML**
A	50	50	50	50	100
B	50	50	0	50	100
C	50	50	0	100	100
D	0	50	0	50	0

B. Variation within data sets (among analyses)			
Area of incongruence	***Consensus***	***Ambiguity***	***Alleles***
A	60	33	20
B	40	33	20
C	20	33	40
D	20	17	40

three of four areas (A, C, and D). The MP-EST tree and STAR tree each had high support for the topology it resolved for two areas (A and D), and the SVDQuartets tree did not have high support for the topology it resolved in any high-conflict area.

For the *ambiguity* data set, the STAR tree had high support for the arrangements it resolved in all four areas. The MP-EST and each ASTRAL trees had high support for the topology it resolved for two areas (A and C), and each SVDQuartets tree had high support for the topology it resolved in one area (D).

For the *alleles* data set, the STAR tree had high support for the arrangements it resolved in all four areas. The MP-EST tree, each ASTRAL tree, and the SVDQuartets tree estimated from phased allele sequences each had high support for the topology it resolved for two areas (MP-EST: A and B; ASTRAL: A and C; SVDQ: C and D), and the SVDQuartets tree estimated from unphased allele sequences had high support for the topology it resolved for one area (D).

Phylogenetic analysis of genes with high levels of deep coalescence

Our data set includes only one individual per species; therefore, non-monophyletic intra-individual alleles are, by default, a case of deep coalescence. To assess whether using allele sequences rather than ambiguity or consensus sequences improves resolution of species trees estimated from gene trees in which deep coalescence is not rare, we performed species tree estimation using ASTRAL on a set of seven gene trees in which more than one in five heterozygous individuals had alleles that were not sister in the gene tree. We did not observe better resolution of relationships in the ASTRAL species tree estimated from the alleles data set gene trees than in the ASTRAL trees estimated from ambiguity and consensus data set gene trees (Appendix S2.20–S2.22). The total percent resolution of *Artocarpus* in the ASTRAL species trees estimated from *consensus*, *ambiguity*, and *alleles* assembly data sets was 68% in each tree. Percent resolution in the CAML trees reconstructed from the consensus and ambiguity assembly data sets was 77% for both trees (Appendix S2.23–S2.24; see methods for explanation of "percent resolution").

DISCUSSION

Our understanding of how deep coalescence of alleles and/or high levels of heterozygosity may affect phylogeny reconstruction is very limited. Phylogeneticists do not always assess heterozygosity in data sets and are usually not aware of deep coalescence of alleles because allele sequences are not assembled or analyzed. In this study, we quantified the amount of deep coalescence of alleles and highly heterozygous sequences and asked how these characteristics of a data set affect our ability to resolve a phylogeny. To answer this question we considered: (1) Do gene trees estimated from genes with high levels of deep coalescence of alleles or heterozygosity have lower resolution than others? (2) Do highly heterozygous individuals and/or those with noncoalescing alleles in gene trees appear to limit resolution in species trees/CAML trees? (3) Does using allele sequences instead of ambiguity or consensus sequences resolve any problems that may be caused by deep coalescence or high heterozygosity? (4) Do phylogenies estimated from allele sequences differ from those estimated from "super-contigs" (consensus or ambiguity-coded contigs)? In our reconstruction of gene trees, CAML trees, and species trees for *Artocarpus* using sequences assembled with three different methods of data assembly, we did not find evidence that the answer to any of these questions is yes.

Highly heterozygous genes and individuals do not appear disproportionately responsible for low resolution or topological uncertainty in gene trees or species/CAML trees

Our finding that highly heterozygous individuals did not influence gene tree resolution or disproportionately occur in high-conflict areas of species or CAML trees was somewhat surprising. The absence of a strong influence of heterozygosity on gene tree resolution was particularly surprising for the *ambiguity* data set, as ambiguity-coded heterozygous positions have previously been shown to decrease phylogenetic resolution (Potts et al., 2014). However, many factors influence phylogenetic resolution, and we cannot isolate the effect of ambiguity-codes or deep coalescence of alleles using our empirical data. In particular, more variable genes are expected to be more phylogenetically informative (Small et al., 2004; Duarte et al., 2010), and high variability among individuals' sequences increases phylogenetic resolution (Parks et al., 2009). Both of these desirable characteristics are inextricably linked to the presence of ambiguity-codes (in ambiguity-coded data) and the increased possibility of non-sisterhood of intra-individual alleles. Our finding that high heterozygosity of individuals or gene trees was also not associated with improvement in phylogenetic resolution is weakly suggestive of these negative effects confounding the positive.

Using allele sequences does not improve phylogenetic resolution or produce meaningfully different topologies in enigmatic areas of the Artocarpus phylogeny

Our ability to answer questions related to "improving" a phylogeny or to comparisons of topologies broadly is limited because we do not know the true phylogeny of real (not simulated) data. We also looked at differences in topologies that resulted from three methods of assembling heterozygous sequences, and used three measures to assess whether any one data set was "better" at resolving the *Artocarpus* phylogeny: (1) bootstrap support in poorly resolved areas of *Artocarpus* evolution, (2) topological consistency across methods of phylogeny reconstruction, and (3) amount of gene tree discordance and gene tree/species tree concordance (i.e, topological consistency across gene trees and between gene trees and species trees). Although we did find that phylogenies reconstructed from the *allelic* sequences differed from those reconstructed with *consensus* or *ambiguity* sequences, topological incongruence among methods used for phylogenetic analysis and between the other two assembly methods was very common. We can only conclude that different methods of data assembly and different programs used for analysis yield different evolutionary hypotheses in poorly supported areas of a phylogeny. This study highlights the uncertainty inherent in estimating the phylogeny of closely related species. We found that high statistical support at several nodes obscured high gene tree incongruence and disagreement among data set assembly and phylogeny reconstruction methods. In particular, CAML analysis of the *consensus* method of data assembly, likely the most commonly used methods of assembly and phylogenetic analysis, consistently resolved topologies with high bootstrap support in high-conflict areas of the *Artocarpus* phylogeny. The use of other data assemblies and other methods of phylogenetic analysis revealed topological uncertainty in these areas,

and in most cases the topology resolved by the consensus-CAML analysis was not the topology resolved by a majority of other analyses. For example, the *alleles* analyses recovered six topologies for the sparsely sampled Area C, none of which were recovered by either CAML analysis, even though the two CAML topologies both had high support (Appendix S3). This incongruence supports previous findings based on simulated and biological data that concatenation analyses can result in the wrong tree with high support under some conditions (Edwards et al., 2007; Kubatko and Degnan, 2007).

In our comparisons of the topologies of, and support for, poorly resolved relationships in *Artocarpus*, no clear patterns emerged to suggest that the use of allele sequences for phylogeny reconstruction improves phylogenetic resolution relative to the use of one sequence per allele. Bootstrap support for high-conflict branches in the tree was not higher in the topologies reconstructed using the alleles data set than in the other topologies. Gene trees estimated from allele sequences were not more concordant than those estimated from the other methods, nor was there more gene tree–species tree concordance for the alleles data set.

Deep coalescence of alleles is sometimes cited as a potential reason for poor resolution toward the tips of a phylogeny (Maddison and Knowles, 2006). Although there is no clear association between the depth of high-conflict branches in the tree and what data set best resolves these areas of the phylogeny, the only high-conflict area for which the topologies resolved by the alleles data set were among the more consistent and better supported topologies was the area that involved branches deeper in the tree than the other three high-conflict areas ("Area A").

Limitations of our results—Our finding that the use of allele sequences does not improve phylogenetic resolution is novel, but its implications are somewhat limited. The extent to which heterozygosity and deep coalescence of alleles affects phylogeny reconstruction will differ for every phylogenetic data set. Our *Artocarpus* data set only included one individual per species; deep coalescence of alleles may be more problematic when resolving relationships among intra-species individuals. However, we have almost certainly under-sampled the allele pool for each species; deep coalescence in *Artocarpus* may be more pervasive than we observe with one individual per species. As described above, we also did not find clear evidence that poor resolution in the *Artocarpus* phylogeny was associated with sequence heterozygosity. If heterozygous sequences are not associated with poor resolution, it is not clear that a more biologically accurate assembly of these genes would improve phylogenetic hypotheses. However, heterozygosity is expected to limit phylogenetic resolution when ambiguity codes are used, as has been shown elsewhere (Kates et al., 2017).

While our method for extracting alleles from sequences obtained through targeted sequencing could be applied to any organism, there are still methodological limitations. Read-backed phasing of coding sequences is restricted by the length of reads and the length of introns. Truly long-range haplotypes (chromosome length) would require long-read sequencing technology or a reference genome. Our method will also struggle with polyploid taxa. Although GATK allows for variant detection at higher ploidy levels, we are not aware of any tools that can create phased haplotypes from target enrichment data.

We used biological data rather than simulated sequences for this study. As such, we do not know the true phylogeny of the study group and cannot determine whether a particular method of sequence assembly yields a more "accurate" phylogeny. Instead, we compared statistical support for relationships resolved and topological consistency across phylogenetic methods for our various data sets. Just as topology depends on the underlying data and the analyses, bootstrap support is also the result of a particular data set and analysis (Soltis and Soltis, 2003); therefore, comparisons of bootstrap support for a topology across data sets or methods are not absolute. For example, the data resampled for bootstrapping in ASTRAL (Mirarab and Warnow, 2015) are bootstrap gene trees, but nucleotide alignment sites are used for bootstrapping by RAxML (Stamatakis, 2014). Furthermore, simulation studies have demonstrated that false positive branches may have inflated multi-locus bootstrap support and that true branches often have low support (Bayzid et al., 2015).

CONCLUSIONS

We present here a novel pipeline for inferring phased alleles from target enrichment data and for systematically evaluating the influence of locus-assembly methods in species tree resolution. While we focused on phylogenetic resolution among species, our method could also be used to extract allelic sequences from targeted sequencing for population genetic analysis within species or for methods of phylogeny reconstruction that use allele frequencies (De Maio et al., 2015). We did not find evidence that the use of allele sequences to reconstruct phylogenies offers a clear improvement over other methods of assembling heterozygous sequences. The diversity of analyses presented here do show the common issue of across-analysis incongruence, which highlights the problems with treating phylogenies as true rather than as evolutionary hypotheses, especially in studies where a single method of sequence assembly and phylogenetic analysis is used. However, the ability to infer phased alleles from target enrichment data presents a number of exciting opportunities to explore whether or not the minimal impact of including alleles in phylogenetic reconstruction that we find is universal, or simply limited to this test case. Testing these methods using clades of varying age and diversity, increasing the sampling of the intraspecific allele pool, and using simulated data will all further our understanding of the questions posed here.

ACKNOWLEDGEMENTS

The authors thank the organizers of this special issue for the invitation to contribute. The authors thank the Field Museum of Natural History for the use of sequencing facilities. This work was funded by National Science Foundation grants to N.J.W. (DEB-1239992 and DEB-1342873) and N.J.C.Z. (DEB-0919119) and a grant to N.J.C.Z. from the Institute for Sustainability and Energy at Northwestern University. The authors also thank two anonymous reviewers and the Associate Editor for providing comments on an earlier version of this manuscript.

DATA ACCESSIBILITY

Illumina sequencing reads have been deposited in the NCBI Sequence Read Archive (BioProject ID PRJNA301299). Multiple sequence alignments for the consensus, ambiguity, and alleles data

sets are available from the Dryad Digital Repository (https://doi.org/10.5061/dryad.r8q72) along with all phylogenetic trees. Scripts used to generate the data sets are freely available at www.github.com/mossmatters/phyloscripts.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

LITERATURE CITED

Ané, C., B. Larget, D. A. Baum, S. D. Smith, and A. Rokas. 2007. Bayesian estimation of concordance among gene trees. *Molecular Biology and Evolution* 24: 412–426.

Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* 19: 455–477.

Bayzid, M. S., S. Mirarab, B. Boussau, and T. Warnow. 2015. Weighted statistical binning: enabling statistically consistent genome-scale phylogenetic analyses. *PLOS One* 10: e0129183.

Berg, C.C., E.J.H. Corner, and F.M. Jarrett. 2006. Moraceae (genera other than *Ficus*). Flora Malesiana, series I, vol. 17, part 1. Nationaal Herbarium Nederland, Leiden, Netherlands.

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

Browning, S. R., and B. L. Browning. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics* 81: 1084–1097.

Capella-Gutiérrez, S., J. M. Silla-Martínez, and T. Gabaldón. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25: 1972–1973.

Charif, D., and J.R. Lobry. 2007. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. *In* U. Bastolla, M. Porto, H. E. Roman, and M. Vendruscolo [eds.], Structural approaches to sequence evolution: molecules, networks, populations, 207–232. Springer-Verlag, Berlin, Germany.

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

De Maio, N., D. Schrempf, and C. Kosiol. 2015. PoMo: An allele frequency-based approach for species tree estimation. *Systematic Biology* 64: 1018–1031.

Drummond, A. J., and A. Rambaut. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* 7: 214.

Duarte, J. M., P. K. Wall, P. P. Edger, L. L. Landherr, H. Ma, J. C. Pires, J. Leebens-Mack, and C. W. dePamphilis. 2010. Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. *BMC Evolutionary Biology* 10: 61.

Edwards, C. E., D. Lefkowitz, D. E. Soltis, and P. S. Soltis. 2008. Phylogeny of *Conradina* and related southeastern scrub mints (Lamiaceae) based on *GapC* gene sequences. *International Journal of Plant Sciences* 169: 579–594.

Edwards, S. V., L. Liu, and D. K. Pearl. 2007. High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences, USA* 104: 5936–5941.

Felsenstein, J. 2003. Inferring phylogenies. Sinauer, Sunderland, MA, USA.

Gardner, E. M., M. G. Johnson, D. Ragone, N. J. Wickett, and N. J. C. Zerega. 2016. Low-coverage, whole-genome sequencing of *Artocarpus camansi* (Moraceae) for phylogenetic marker development and gene discovery. *Applications in Plant Sciences* 4: 1600017.

Heled, J., and A. J. Drummond. 2010. Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution* 27: 570–580.

Huerta-Cepas, J., F. Serra, and P. Bork. 2016. ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Molecular Biology and Evolution* 33: 1635–1638.

Johnson, M. G., E. M. Gardner, Y. Liu, R. Medina, B. Goffinet, A. J. Shaw, N. J. C. Zerega, and N. J. Wickett. 2016. HybPiper: extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in Plant Sciences* 4: 1600016.

Kates, H. R., P. S. Soltis, and D. E. Soltis. 2017. Evolutionary and domestication history of *Cucurbita* (pumpkin and squash) species inferred from 44 nuclear loci. *Molecular Phylogenetics and Evolution* 111: 98–109.

Katoh, K., and D. M. Standley. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* 30: 772–780.

Keller, O., M. Kollmar, M. Stanke, and S. Waack. 2011. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* 27: 757–763.

Knowles, L. L. 2009. Estimating species trees: methods of phylogenetic analysis when there is incongruence across genes. *Systematic Biology* 58: 463–467.

Kubatko, L. S., B. C. Carstens, and L. L. Knowles. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics* 25: 971–973.

Kubatko, L. S., and J. H. Degnan. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology* 56: 17–24.

Li, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*: 1303.3997 [q-bio.GN]. Available at http://arxiv.org/abs/1303.3997.

Liu, L., D. K. Pearl, R. T. Brumfield, and S. V. Edwards. 2008. Estimating species trees using multiple-allele DNA sequence data. *Evolution* 62: 2080–2091.

Liu, L., S. Wu, and L. Yu. 2015. Coalescent methods for estimating species trees from phylogenomic data. *Journal of Systematics and Evolution* 53: 380–390.

Liu, L., L. Yu, and S. V. Edwards. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology* 10: 302.

Liu, L., L. Yu, L. Kubatko, D. K. Pearl, and S. V. Edwards. 2009a. Coalescent methods for estimating phylogenetic trees. *Molecular Phylogenetics and Evolution* 53: 320–328.

Liu, L., L. Yu, D. K. Pearl, and S. V. Edwards. 2009b. Estimating species phylogenies using coalescence times among sequences. *Systematic Biology* 58: 468–477.

Maddison, W. P., and L. L. Knowles. 2006. Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology* 55: 21–30.

Maddison, W. P., and D. R. Maddison. 2017. Mesquite: a modular system for evolutionary analysis. Computer program and documentation distributed by the author. Available at http://mesquiteproject.org.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, et al. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20: 1297–1303.

Mirarab, S., and T. Warnow. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* 31: i44–52.

Paradis, E., J. Claude, and K. Strimmer. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20: 289–290.

Parks, M., R. Cronn, and A. Liston. 2009. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biology* 7: 84.

Patterson, M., T. Marschall, N. Pisanti, L. van Iersel, L. Stougie, G. W. Klau, and A. Schönhuth. 2015. WhatsHap: weighted haplotype assembly for future-generation sequencing reads. *Journal of Computational Biology* 22: 498–509.

Potts, A. J., T. A. Hedderson, and G. W. Grimm. 2014. Constructing phylogenies in the presence of intra-individual site polymorphisms (2ISPs) with a focus on the nuclear ribosomal cistron. *Systematic Biology* 63: 1–16.

Ranwez, V., S. Harispe, F. Delsuc, and E. J. P. Douzery. 2011. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. *PLOS One* 6: e22594.

Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19: 1572–1574.

Rothfels, C. J., K. M. Pryer, and F. W. Li. 2017. Next-generation polyploid phylogenetics: rapid resolution of hybrid polyploid complexes using PacBio single-molecule sequencing. *New Phytologist* 213: 413–429.

Small, R. L., R. C. Cronn, and J. F. Wendel. 2004. Use of nuclear genes for phylogeny reconstruction in plants. *Australian Systematic Botany* 17: 145–170.

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

Soltis, P. S., and D. E. Soltis. 2003. Applying the bootstrap in phylogeny reconstruction. *Statistical Science* 18: 256–267.

Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22: 2688–2690.

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

Swofford, D.L. 2002. PAUP*: phylogenetic analysis using parsimony (*and other methods), version 4.0b10. Sinauer, Sunderland, MA, USA.

Uribe-Convers, S., M. L. Settles, and D. C. Tank. 2016. A phylogenomic approach based on PCR target enrichment and high throughput sequencing: resolving the diversity within the South American species of *Bartsia* L. (Orobanchaceae). *PLOS One* 11: e0148203.

Waskom, M., O. Botvinnik, P. Hobson, J. Warmenhoven, J.B. Cole, Y. Halchenko, J. Vanderplas, et al. 2014. Seaborn: statistical data visualization. https://doi.org/10.5281/zenodo.12710

Weisrock, D. W., S. D. Smith, L. M. Chan, K. Biebouw, P. M. Kappeler, and A. D. Yoder. 2012. Concatenation and concordance in the reconstruction of mouse lemur phylogeny: an empirical demonstration of the effect of allele sampling in phylogenetics. *Molecular Biology and Evolution* 29: 1615–1630.

Williams, E. W., E. M. Gardner, R. Harris 3rd, A. Chaveerach, J. T. Pereira, and N. J. C. Zerega. 2017. Out of Borneo: biogeography, phylogeny and divergence date estimates of *Artocarpus* (Moraceae). *Annals of Botany* 119: 611–627.

Zerega, N. J. C., M. N. N. Supardi, and T. J. Motley. 2010. Phylogeny and recircumscription of Artocarpeae (Moraceae) with a focus on *Artocarpus*. *Systematic Botany* 35: 766–782.