Paralogs and off-target sequences improve phylogenetic resolution in a densely-sampled study of the breadfruit genus (*Artocarpus*, Moraceae)

Elliot M. Gardner[1,2,3,4,a,b], Matthew G. Johnson[5,a], Joan T. Pereira[6], Aida Shafreena Ahmad Puad[7], Deby Arifiani[8], Sahromi[9], Norman J. Wickett[1], and Nyree J.C. Zerega[1,2,b]

[1] Chicago Botanic Garden, Negaunee Institute for Plant Conservation Science and Action, 1000 Lake Cook Road, Glencoe, IL, 60022, USA

[2] Northwestern University, Plant Biology and Conservation Program, 2205 Tech Dr., Evanston, IL, 60208, USA

[3] The Morton Arboretum, 4100 IL-53, Lisle, Illinois 60532, USA (current affiliation)

[4] Case Western Reserve University, Department of Biology, 2080 Adelbert Road, Cleveland, Ohio 44106, USA (current affiliation)

[5] Texas Tech University, Department of Biological Sciences, 2901 Main Street, Lubbock, TX, 79409-3131, USA

[6] Forest Research Centre, Sabah Forestry Department, P.O. Box 1407, 90715 Sandakan, Sabah, Malaysia

[7] Universiti Malaysia Sarawak, Kuching, Sarawak, Malaysia

[8] Herbarium Bogoriense, Research Center for Biology, Indonesian Institute of Sciences, Cibinong, Jawa Barat, Indonesia

[9] Center for Plant Conservation Botanic Gardens, Indonesian Institute Of Sciences, Bogor, Jawa Barat, Indonesia

[a] Co-first authors

[b] For correspondence, email elliot.gardner@case.edu or n-zerega@northwestern.edu

Abstract

We present a 517-gene phylogenetic framework for the breadfruit genus *Artocarpus* (ca. 70 spp., Moraceae), making use of silica-dried leaves from recent fieldwork and herbarium specimens (some up to 106 years old) to achieve 96% taxon sampling. We explore issues relating to assembly, paralogous loci, partitions, and analysis method to reconstruct a phylogeny that is robust to variation in data and available tools. While codon partitioning did not result in any substantial topological differences, the inclusion of flanking non-coding sequence in analyses significantly increased the resolution of gene trees. We also found that increasing the size of datasets increased convergence between analysis methods but did not reduce gene tree conflict. We optimized the HybPiper targeted-enrichment sequence assembly pipeline for short sequences derived from degraded DNA extracted from museum specimens. While the subgenera of *Artocarpus* were monophyletic, revision is required at finer scales, particularly with respect to widespread species. We expect our results to provide a basis for further studies in *Artocarpus* and provide guidelines for future analyses of datasets based on target enrichment data, particularly those using sequences from both fresh and museum material, counseling careful attention to the potential of off-target sequences to improve resolution.

Key words: Phylogenomics, target enrichment, non-coding sequences, Moraceae, *Artocarpus*

Despite the increasing availability of whole-genome sequencing, costs and computational limitations still make it impractical, and often unnecessary, for large phylogenetic projects. Reduced-representation methods such as target enrichment sequencing (e.g., HybSeq) have become important tools for phylogenetic studies, enabling high-throughput and cost-effective sequencing of hundreds of informative loci (Faircloth et al. 2012; Mandel et al. 2014; Weitemier et al. 2014). HybSeq involves hybridizing a randomly-sheared sequencing library (e.g., Illumina) to oligonucleotide bait sequences, typically exonic sequences from one or more taxa within or near the target clade. The resulting sequence data include exons and flanking non-coding sequences (e.g., introns, UTRs). While HybSeq recovers fewer loci than the tens of thousands available from RAD-seq, it is more repeatable, recovers longer loci, and has much lower rates of missing data (Weitemier et al. 2014). Accordingly, researchers have successfully employed HybSeq in studies ranging from deep phylogenetics (Prum et al. 2015; Liu et al. 2019) to within-species phylogeography (Villaverde et al. 2018). Like many high throughput sequencing methods, including transcriptomics and RAD-seq, HybSeq can provide hundreds of thousands of characters for phylogenetic reconstruction. Making the most of these large datasets requires careful attention to assembly and analysis methods, particularly when dealing with degraded DNA extracted from museum specimens.

HybSeq can be used to recover sequences from degraded DNA extracted from old museum specimens because it relies on capturing short fragments of DNA using 120 bp probes, rather than two flanking primer sequences as in amplicon sequencing. Therefore, HybSeq may succeed where direct PCR-based methods might fail (Staats et al. 2013), substantially raising the value of natural history collections for phylogenetic studies and offering the possibility of including taxa that are not possible to collect due to extinction or infeasible fieldwork (Buerki

and Baker 2016; Brewer et al. 2019). Extreme fragmentation of old DNA and contamination with non-endogenous sequences, such as fungi, still present challenges to assembly of sequences from short-read platforms, leading many studies to focus on high-copy regions of DNA, such as chloroplast sequences (Bakker et al. 2016). Enriching for targeted genes can transform low-copy nuclear genes into high-copy components of a sequencing library, making HybSeq particularly appropriate for museum or herbarium material (Hart et al., 2016; Villaverde et al., 2018).

Ensuring that all sequences used for phylogenetic reconstruction are orthologous is a key step in any phylogenetics workflow. HybSeq will recover any portions of the genome that are sufficiently similar to the bait sequences—up to 25–30% divergence in some cases (Johnson et al. 2019; Liu et al. 2019)—including paralogs or genes with enough shared domains (Hart et al. 2016; Johnson et al. 2016). For this reason, HybSeq bait development has usually focused on single to low-copy genes (Chamala et al. 2015; Gardner et al. 2016). Nevertheless, whole genome duplications can render single copy genes double copy in entire clades. We previously developed HybSeq baits for phylogenetic reconstruction of the genus *Artocarpus* J.R. Forst. & G. Forst. (Moraceae) from a three-way orthology search between closely related species in the Rosales: *Cannabis sativa* L. (Cannabaceae) (van Bakel et al. 2011), *Morus notabilis* C.K. Schneid. (Moraceae) (He et al. 2013), and *Artocarpus camansi* Blanco (Gardner et al. 2016). Due to an ancient whole-genome duplication in *Artocarpus*, many of the 333 genes were represented as paralogous pairs in that genus, although in almost all cases they were diverged enough to sort and analyze separately (Johnson et al. 2016). The impact of this by-catch on phylogenetic reconstruction remains unclear, but has the potential to greatly increase the number of phylogenetically informative genes, if the paralogs can be easily sorted.

Accurate phylogenetic reconstruction relies on carefully considered parameters for sequence alignment, trimming, and partitioning of the character matrix, among other factors. Because HybSeq recovers not only targeted exons but also flanking non-coding sequences, contigs containing both exons and introns are often aligned and analyzed together (e.g. (Medina et al. 2019). This can make it difficult to ensure that exons are aligned in frame, particularly when frameshifts are present due to the presence of pseudogenes, sequencing errors, or other processes (Ranwez 2011). Aligning coding and noncoding regions together can also hamper partitioning of datasets by codon position, but how these issues impact phylogenetic reconstruction remains unclear (Xi et al. 2012; Lanfear et al. 2014).

Concatenating all loci into a supermatrix can result in near-perfect bootstrap support for almost all nodes in a phylogeny (Sayyari and Mirarab 2016). Despite this apparent high support, there may be substantial discordance among gene histories due to incomplete lineage sorting (Kubatko and Degnan 2007; Degnan and Rosenberg 2009). Although there is an increased availability of efficient methods that are consistent under the predictions of the multi-species coalescent model, clear results can be obscured if the underlying gene trees are uninformative (Smith et al. 2015; Sayyari et al. 2017). A major advantage of HybSeq over methods with short, anonymous loci, or large amounts of missing data, is that loci are both long enough to generate single-gene phylogenies and subject to few enough missing taxa per locus for those single-gene phylogenies to be informative. In this paper we explore the possibilities of HybSeq, including the informativeness of paralogs, effects of missing data, and utility of herbarium specimens to reconstruct the most data-rich phylogeny of the breadfruit genus to date.

Study system

Artocarpus J.R. Forst. & G. Forst. (Moraceae, Figure 1) contains approximately 70 species of monoecious trees with a center of diversity in Borneo and a native range that extends from India to the Solomon Islands (Williams et al. 2017). The genus is best known for important but underutilized crops such as breadfruit (*A. altilis* (Parkinson) Fosberg) and jackfruit (*A. heterophyllus* Lam.), and it also contains crops of regional importance like cempedak (*A. integer* (Thunb.) Merr.) and tarap (*A. odoratissimus* Blanco), and more than a dozen other species with edible fruits whose potential remains largely unexplored (Zerega et al. 2010, 2015; Wang et al. 2018; Witherup et al. 2019).

Artocarpus is characterized by spicate to globose staminate ("male") inflorescences composed of tiny flowers bearing one stamen each. Carpellate ("female") inflorescences are composed of tightly packed tiny flowers. In most cases, adjacent carpellate flowers are at least partially fused together. Carpellate inflorescences develop into syncarps, which are tightly packed accessory fruits composed mainly of fleshy floral tissue. Syncarps of different species range in size from a few centimeters in diameter to over half a meter long. *Artocarpus* is the largest genus in the tribe Artocarpeae, which also includes two smaller Neotropical genera, *Batocarpus* H. Karst. (3 spp.) and *Clarisia* Ruiz & Pav. (3 spp.). The neotropical genera always have spicate staminate inflorescences; carpellate flowers may be solitary or condensed into globose heads, but neither tepals nor adjacent flowers are fused.

The most recent complete revision of *Artocarpus* (Jarrett 1959, 1960) recognized two subgenera, *Artocarpus* and *Pseudojaca* Tréc., distinguished by phyllotaxy (leaf arrangement), and the degree of fusion between adjacent carpellate flowers. Since then, several new species have been described by Jarrett and others (Jarrett, 1975; Zhengyi and Xiushi, 1989; Kochummen, 1998; Berg, 2005; Gardner et al. in prep.; Gardner and Zerega in prep.). Berg et al. (2006)

136 revised the Malesian species for the *Flora Malesiana*, in a few cases combining several taxa into
137 a broadly-circumscribed single species. Examples include *A. altilis* (encompassing *A. altilis, A.*
138 *camansi, A. mariannensis* Tréc., *A. horridus* F.M. Jarrett, *A. blancoi* Merr., *A. pinnatisectus*
139 Merr., and *A. multifidus* F.M. Jarrett) and *A. lacucha* Buch.-Ham. (encompassing *A. lacucha, A.*
140 *dadah* Miq., *A. fretessii* Teijsm. & Binnend., *A. ovatus* Blanco, and *A. vrieseanus* var. *refractus*
141 Becc. (F.M. Jarrett)). In these cases, this paper follows Jarrett's nomenclature for clarity. The
142 most recent circumscription of *Artocarpus* recognized four subgenera within *Artocarpus* (Figure
143 1) and was based on just two gene regions and approximately 50% of taxa (Zerega et al. 2010).
144 The subgenera were distinguished by phyllotaxy, the degree of fusion between adjacent
145 carpellate flowers, and the position of inflorescences on the tree (axillary or cauliflorous). A
146 well-sampled phylogenetic framework for *Artocarpus* is necessary to inform future taxonomic
147 revision and to clarify relationships within this important genus, in particular the relationships
148 between crop species and their wild relatives, whose conservation is a priority (Castañeda-
149 Álvarez et al. 2016).

150 In this study, we used near-complete (80/83) taxon sampling (at the subspecies level or
151 above) in *Artocarpus* to explore the impact of paralogs, codon partitions, noncoding sequences,
152 and analysis method (species tree versus concatenated supermatrix) on phylogenetic
153 reconstruction in order to develop best practices for the analysis of HybSeq data and to explore
154 the limits of phylogenomic resolution. We also used this data set to improve the target capture
155 assembly pipeline HybPiper, which is now optimized for accurately scaffolding small
156 disconnected contigs resulting from degraded DNA. The objectives of the study were to (1) Use
157 broad sampling from silica dried material and herbarium specimens over 100 years old to
158 achieve near complete taxon sampling for *Artocarpus*; (2) examine the impact of paralogs,

partitions, and analysis method on phylogenetic reconstruction; and (3) test the monophyly of the current taxonomic divisions within *Artocarpus* to provide a phylogenetic framework for future studies on the taxonomy, conservation, and ecology of the genus.

MATERIALS AND METHODS

Taxon sampling

We sampled all *Artocarpus* taxa at the subspecies level or above recognized by Jarrett (1959, 1960), Berg et al. (2006), and Kochummen (1998), all three obsolete species that Jarrett (1959) sunk into *A. treculianus* Elmer, and all of the new species described by Wu and Zhang (1989), for a total of 83 named *Artocarpus* taxa. We also sampled nine taxa of questionable affinities. We replicated samples across geographic or morphological ranges when possible, for a total of 167 ingroup samples. As outgroups, we sampled one member of each genus in the Neotropical Artocarpeae (*Batocarpus* and *Clarisia*) and the sister tribe Moreae (*Morus* L., *Streblus* Lour., *Milicia* Sim., *Trophis* P. Browne, *Bagassa* Aubl., and *Sorocea* A. St.-Hil.). We obtained samples from our own field collections preserved in silica gel (from Malaysia, Thailand, Hong Kong, Bangladesh, and India, and from botanic gardens in Indonesia, Malaysia, and Hawai'i, USA) and from herbarium specimens up to 106 years old (from the following herbaria: BM, BO, CHIC, E, F, HAST, HK, K, KUN, L, MO, NY, KEP, S, SAN, SNP, US). In total we included 179 samples (Table S1).

Sample preparation and sequencing

We sampled approximately 0.5 cm_2 of dried leaf from each sample for DNA extraction. For herbarium specimens, we sampled from a fragment packet when feasible and when it was clear that the material in the fragment packet originated from the specimen on the sheet (something that cannot always be assumed with very old specimens). DNA was extracted using one of three methods; (1) the Qiagen DNeasy Plant Mini Kit (Qiagen, Valencia, California, USA) following the manufacturer's protocol; (2) the MoBio PowerPlant Pro DNA Kit, (MoBio Laboratories, Carlsbad, California, USA); or (3) a modified CTAB protocol (Doyle and Doyle 1987). For kit extractions, the protocols were modified for herbarium material by extending initial incubation times (Williams et al., 2017) and adding an additional 200 µL of ethanol to the column-binding step. CTAB extractions of herbarium specimens, which often had high but impure DNA yields, were cleaned using a 1:1.8:5 ratio of sample, SPRI beads, and isopropanol, the latter added to prevent the loss of small fragments (Lee 2014). For herbarium specimens, we sometimes combined two or more separate extractions in order to accumulate enough DNA for library preparation. We assessed degradation of DNA from herbarium specimens using either an agarose gel or a High-Sensitivity DNA Assay on a BioAnalyzer 2100 (Agilent) and did not sonicate samples whose average fragment size was less than 500bp. The remaining DNA samples were sonicated to a mean insert size of 550bp using a Covaris M220 (Covaris, Wobum, Massachussetts, USA). Libraries were prepared with either the Illumina TruSeq Nano HT DNA Library Preparation Kit (Illumina, San Diego, California, USA) or the KAPA Hyper Prep DNA Library Kit following the manufacturer's protocol, except that reactions were performed in one-third volumes to save reagent costs. We used 200ng of input DNA when possible; for some samples, input was as low as 10ng. For herbarium samples with degraded DNA, we usually did not perform size selection, unless there were some fragments that were above 550bp. We also

diluted the adapters from 15 µM to 7.5 µM, and usually performed only a single SPRI bead cleanup between adapter ligation and PCR amplification. Many of these libraries contained substantial amounts of adapter dimer, so we adjusted the post-PCR SPRI bead cleanup ratio to 0.8x. Libraries were enriched for 333 phylogenetic markers (Gardner et al., 2016) with a MYbaits kit (MYcroarray, Ann Arbor, Michigan, USA) following the MYbaits manufacturer's protocol (version 3). Hybridization took place in pools of 6–24 libraries; within each pool, we used equal amounts of all libraries (20–100ng, as available), and tried to avoid pooling samples with dramatically different phylogenetic distances to the bait sequnces (*Morus* and *Artocarpus*), as closer taxa can out-compete multiplexed distant taxa in hybridization reactions, as we previously found when pooling *Dorstenia* L. and *Parartocarpus* Baill. with *Artocarpus* (Johnson et al. 2016). We reamplified enriched libraries with 14 PCR cycles using the conditions specified in the manufacturer's protocol. In some cases, adapter dimer remained even after hybridization; in those cases, we removed it either using a 0.7x SPRI bead cleanup or, in cases where the library fragments were very short (ca. 200bp, compared to 144bp for the dimer), by size-selecting the final pools to >180bp on a BluePippin size-selector using a 2% agarose gel cassette (Sage Science, Beverley, Massachussetts, USA). Pools of enriched libraries were sequenced on an Illumina MiSeq (600 cycle, version 3 chemistry) alongside samples for other studies in three multiplexed runs each containing 30–99 samples.

Sequence quality control and analyses

Demultiplexing and adapter trimming took place automatically through Illumina BaseSpace (basespace.illumina.com). All reads have been deposited in GenBank (BioProject no. PRJNA322184). Raw reads were quality trimmed using Trimmomatic (Bolger et al., 2014), with

a quality cutoff of 20 in a 4-bp sliding window, discarding any reads trimmed to under 30 bp. In addition to the samples sequenced for this study, reads used for assemblies included all *Artocarpus* samples sequenced in Johnson et al. (2016) (available under the same BioProject number). Common methods for target capture assembly include mapping reads to a reference (Weitemier et al. 2014; Hart et al. 2016) and *de novo* assemblies (Mandel et al. 2014; Faircloth 2015), but both have drawbacks. Read mapping can result in lost data, particularly indels and non-coding regions, unless a close reference is available. On the other hand, *de novo* assemblies can also result in lost data if loci cannot be assembled into single scaffolds. A compromise approach, implemented in HybPiper, is to combine local *de novo* assemblies—which may result in many small contigs per locus—with scaffolding based on a reference coding sequence, which need not be closely related; a reference with less than 30% sequence, typically within the same family or order, will usually suffice (Johnson et al. 2016, 2019). The resulting assemblies thus cover the maximum available portion of each locus, notwithstanding the existence of long gaps, and also make use of all available on-target reads, including introns, not simply those that can be aligned to a reference.

We assembled sequences using HybPiper 1.2, which represented an update of the original pipeline optimized for short reads from highly-fragmented DNA from museum specimens. HybPiper's guided assembly method uses the reference to scaffold localized de novo assemblies. This is particularly advantageous when dealing with very short reads from degraded DNA, because for those samples, reads covering a single exon may assemble into more than one contig. In those cases, HybPiper uses the reference to scaffold and concatenate multiple contigs into a "supercontig" containing the gene of interest as well as any flanking noncoding sequences (Johnson et al. 2016). The new version of HybPiper is optimized to accurately handle many

small contigs covering a single gene, deduplicating overlaps and outputting high-confidence predicted coding sequences even in the presence of many gaps caused by fragmentary local assemblies. HybPiper as well as all related scripts used in this study are available at https://github.com/mossmatters/HybPiper and https://github.com/mossmatters/phyloscripts. We generated a new HybPiper reference for this study, using reads from all four subgenera of *Artocarpus*. Target-enriched reads from *A. camansi* Blanco (the same individual used for whole-genome sequencing in the original marker development (Gardner et al. 2016), *A. limpato* Miq., *A. heterophyllus*, and *A. lacucha* (the latter three from reads sequenced in Johnson 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. Predicted genes were annotated using a BLASTn search seeded with the HybPiper target file of 333 phylogenetic marker genes from Johnson et al. (2016). Paralogs were annotated as follows: genes covering at least 75% of the primary ortholog (labeled "p0") were labeled as "paralogs" ("p1", "p2", etc.). Genes covering less than 75% of the primary ortholog (labeled "e0") were labeled as "extras" ("e1", "e2", etc.), denoting uncertainty as to whether they are paralogs or merely genes with a shared domain. Single copy genes were labeled as "single" in the new reference. We used this new 4-taxon reference to guide all ingroup assemblies, and we used the original set of *Morus notabilis* targets (Johnson et al., 2016) to guide all outgroup assemblies.

We set the per-gene coverage cutoff to 8x, except for certain low-read samples where gene recovery was improved by lowering the coverage cutoff to 4x (10 samples) or 2x (18 samples). HybPiper relies on SPAdes for local *de novo* assemblies. SPAdes creates several assemblies with different k-mer values, with the maximum estimated from the reads (up to 127bp), and then merges them into a final assembly. For herbarium samples that initially

recovered fewer than 400 genes, we reran HybPiper, manually setting the maximum k-mer values for assembly to 55 instead of allowing SPAdes to automatically set it. To extract non-coding sequences and annotate gene features along assembled contigs, we used the HybPiper script "intronerate.py". We assessed target recovery success using the get_seq_lengths.py and gene_recovery_heatmap.r scripts from HybPiper.

To mask low-coverage regions likely to contain sequencing errors, we mapped each sample's reads to its HybPiper supercontigs using BWA (Li and Durbin 2009), removed PCR duplicates using Picard (Broad Institute 2016), and calculated the depth at each position with Samtools (Li et al. 2009). Using BedTools (Quinlan and Hall 2010), we then hard-masked all positions covered by less than two unique reads. We then used the masked supercontigs and the HybPiper gene annotation files to generate masked versions of the standard HybPiper outputs (using intron_exon_extractor.py): (1) the predicted coding sequence for each target gene ("exon"); (2) the entire contig assembled for each gene ("supercontig"); and (3) the predicted non-coding sequences for each gene ("non-coding", including introns, UTRs, and intergenic sequences).

To the HybPiper output, we added the original orthologs (CDS only) identified in *Morus notabilis* (Gardner et al., 2016). Because paralogs were only assembled for ingroup samples (due to an *Artocarpus*-specific whole-genome duplication (Gardner et al. 2016)), we added the corresponding "p0" or "e0" from *Morus* to each paralog alignment to serve as an outgroup.

We filtered each set of sequences as follows. For "exon" sequences, we subtracted masked bases (Ns) and removed sequences less than 150 bp and sequences covering less than 20% of the average sequence length for that gene. For "supercontig" sequences, we removed

295 sequences whose corresponding "exon" sequences had been removed. Samples with less than
296 100 genes remaining after filtering were excluded from the main analyses.
297 Alignment and trimming then proceeded as follows. For "exon" output, after removing
298 the genes and sequences identified during the filtering stage, we created in-frame alignments
299 using MACSE (Ranwez 2011). For "supercontig" output, we used MAFFT for alignment (--
300 maxiter 1000) (Katoh and Standley 2013). We trimmed all alignments to remove all columns
301 with >75% gaps using Trimal (Capella-Gutiérrez et al. 2009).
302 To quickly inspect gene trees for artifacts, we built gene trees from the trimmed "exon"
303 alignments using FastTree (Price et al. 2009) and visually inspected the gene trees for outlier
304 long branches within the ingroup to identify alignments containing improperly sorted paralogous
305 sequences. In some cases, we visually inspected alignments using AliView (Larsson 2014). We
306 discarded a small number of genes whose alignments contained paralogous sequences, for a final
307 set of 517 genes, including all of the original 333 genes.
308 We used the trimmed alignments to create three sets of gene alignment datasets:
309 1. *CDS:* "exon" alignments, not partitioned by codon position;
310 2. *Partitioned CDS*: 333 "exon" alignments, partitioned by codon position; and
311 3. *Supercontig:* "supercontig" alignments, not partitioned within genes
312 We also attempted to create a codon-partitioned supercontig alignment by separately
313 aligning "exon" and "intron" sequences and then concatenating them, resulting in three partitions
314 per gene. However, this dataset differed substantially from the *supercontig* dataset, resulting in
315 substantially differing (and nonsensical) topologies even when the partitions were removed;
316 samples with a high proportion of very short or missing non-coding sequences clustered together,
317 perhaps because aligning very short non-coding sequences without longer coding sequences to

anchor them produced unreliable alignments. We therefore did not include the *partitioned exon+intron* dataset in the main analyses (discussed further in Appendix 1).

To investigate whether including both copies of a paralogous locus impacted phylogenetic reconstruction, we created versions of each dataset with and without paralogs. We analyzed each of these six datasets using the following two methods, for a total of 12 analyses: (A) *Concatenated supermatrix:* all genes were concatenated into a supermatrix, with each gene partitioned separately (i.e. 1 or 3 partitions per gene, depending on the dataset) and analyzed using RAxML 10 (Stamatakis, 2006) under GTR+CAT model with 200 rapid bootstrap replicates, rooted with the Moreae outgroups; (B) *Species tree:* each gene alignment was analyzed using RAxML 10 under the GTR+CAT model with 200 rapid bootstrap replicates, rooted with the Moreae outgroups. Nodes with <33% support were collapsed into polytomies using SumTrees (Sukumaran and Holder 2010), and the resulting trees were used to estimate a species tree with ASTRAL-III (Mirarab and Warnow, 2015). We estimated node support with multilocus bootstrapping (-r, 160 bootstrap replicates) and by calculating the proportion of quartet trees that support each node (-t 1) (Mirarab and Warnow 2015; Zhang et al. 2017). For the final trees, we also used SumTrees to calculate the proportion of gene trees supporting each split. Quartet support is directly related to the method ASTRAL uses for estimating species trees—decomposing gene trees into quartets (Mirarab and Warnow 2015); it is also less sensitive to occasional out-of-place taxa than raw gene-tree support.

Because all RAxML analyses were conducted using the GTRCAT model, we also repeated the analyses of the CDS datasets using the GTRGAMMA model to investigate the robustness of the recovered topologies to slight model differences.

To summarize the overall bootstrap support of each tree with a single statistic, we calculated "percent resolution" as the number of bipartitions with >50% bootstrap support divided by the total number of bipartitions and represents the proportion of nodes that one might consider resolved (Kates et al. 2018). We visualized trees using FigTree (Rambaut 2016) and the APE package in R (Paradis et al. 2004). To compare trees, we used the phytools package in R (Revell 2012) to plot a consensus tree and to calculate a Robison-Foulds (RF) distance matrix for all trees. The RF distance between tree *A* and tree *B* equals the number of bipartitions unique to *A* plus the number of bipartitions unique to *B*. We visualized the first two principal components of the matrix using the Lattice package in R (Sarkar 2008). In addition, we conducted pairwise topology comparisons using the "phylo.diff" function from the Phangorn package in R (Schliep 2011) and an updated version of "cophylo" from phytools (github.com/liamrevell/phytools/). All statistical analyses took place in R (R Core Development Team, 2008).

Supermatrix analyses took place on the CIPRES Science Gateway (Miller et al. 2010). All other analyses took place on a computing cluster at the Chicago Botanic Garden, and almost all processes were run in parallel using GNU Parallel (Tange 2018). Alignments and trees have been deposited in the Dryad Data Repository (accession no. TBA).

RESULTS

Sequencing and assembly

Of the 179 sequenced accessions, 164 resulted in successful HybPiper assemblies (>25 genes) (Figure 2, Table S1), including all attempted taxa except for *A. scandens* Miq. sensu

Jarrett—included by Berg et al. (2006) in *A. scandens*, which was assembled—*A. nigrifolius* C.Y. Wu and *A. nanchuanensis*, C.Y. Wu, two species closely allied to *A. hypargyreus* Hance, which was also assembled. The less successful samples almost all had very few reads and may have been out-competed by other samples during hybridization, reamplification, or both. Fewer reads were also associated with shorter assembled sequences (Figure 2). Five out of the 164 HybPiper assemblies included sequences for less than 100 genes after filtering and were excluded from the main analyses. This included the only remaining accession of *A. reticulatus* Miq. Adding the *Morus notabilis* sequences therefore resulted in a main data set containing 160 samples representing 80 out of 83 named *Artocarpus* taxa at the subspecies/variety level or above (96%), in addition to nine taxa of uncertain affinity.

Overall, samples collected more recently showed improved sequencing results (Figure 2), primarily because the majority of samples collected since 2000 were dried on silica gel. Whether a sample was dried on silica gel was significantly associated with increased gene length as a percentage of average length ($R^2 = 0.33$, $P < 0.0001$) and to a lesser extent with the total number of genes recovered ($R^2 = 0.17$, $P < 0.0001$). All 16 unsuccessful (<25 genes) assemblies were taken from herbarium sheets (with collection years spanning 1917 to 1997), rather than from silica-dried material. Among 67 successfully-assembled samples taken from herbarium sheets, younger age was significantly associated only with increased gene length, although the model was a poor fit ($R^2 = 0.06$, $P = 0.02728$), and not with an increase in the number of genes recovered ($P = 0.2833$) (Figure 3). By the same token, we observed a decrease in average DNA fragment size in older samples (Figure S1). Likewise, lowering the maximum assembly k-mer values for herbarium samples with under 400 genes increased recovery by an average of 20 genes.

386 Gene recovery was high; the average sample (of the 160 passing the final filter) had
387 sequences for 448/517 genes (87%). In the final filtered dataset of 333 genes, the average gene
388 had exon sequences for 151/160 samples (94%, range 57–160, median 154) and non-coding
389 sequences for 130 (81%, range 49–151, median 131). For the 184 paralogs, the average gene in
390 the final filtered dataset had exon sequences for 117/160 samples (73%, range 32–148, median
391 126) and intron sequences for 101 (63%, range 30–132, median 110) (Table S2). The
392 supermatrix of trimmed "exon" alignments for the primary 333 genes contained 407,310
393 characters; and the full set of 517 "exon" alignments, including 184 paralogs, contained 569,796
394 characters. The supermatrix of 333 trimmed "supercontig' alignments contained 813,504
395 characters, and the full set of 517 genes contained 1,181,279 characters. The full set of "exon"
396 alignments had 21% gaps or undetermined characters, while the full set of "supercontig"
397 alignments was 36.87% gaps or undetermined characters.
398
399 *Phylogenetic disagreement*
400 A strict consensus of the 12 species trees under GTR+CAT, with a 20% length cutoff
401 (henceforth, the "main analysis") had 100/159 (63%) nodes resolved (mean RF distance 53),
402 revealing complete agreement among the various analyses in backbone relationships but
403 substantial disagreement at shallower nodes (Figure 5). The six ASTRAL phylogenies differed
404 little from one another, whereas the supermatrix analyses had somewhat greater divergence
405 (Figure 4, Figure S2).
406 *Partitions and model selection* — In the exon datasets, partitioning by codon position
407 (Figures 4, S2, S3) had little impact on the final topology, resulting in a single rearrangement
408 within *A. lacucha* s.s. in the supermatrix analysis (RF 4), and in the ASTRAL analysis a change

409 in the position of an undetermined sample and the change of *A. sepicanus* + *A. altissimus* from a
410 clade to a grade (RF 12). The choice of model (GTRCAT vs GTRGAMMA) also produced only
411 minute changes (Figures S2, S7).

412 *Paralogs*— The addition of paralogs led to slightly more disagreement (Figures 4, S4,
413 Table S3). In the exon dataset, changes to the positions of *A. nitidus* ssp. *lingnanensis* (Merr.)
414 F.M. Jarrett and *A. gomezianus* Wall. ex Tréc. affected the backbone of ser. *Peltati* F.M. Jarrett,
415 subg. *Pseudojaca*, in the supermatrix analysis (RF 58); in the ASTRAL analysis, there were
416 somewhat fewer rearrangements, but the mainly occurred in the same clade (RF 20). However,
417 the disagreement was reduced when noncoding sequences were included (supermatrix RF 22;
418 ASTRAL RF 8).

419 *Introns*— The inclusion of non-coding sequences (Figures 4, S5, Table S3) led to similar
420 amounts of disagreement, with rearrangements at the series level in subg. *Pseudojaca* and subg.
421 *Artocarpus*. Disagreement was greater in the supermatrix analyses (no paralogs) (RF 62) than in
422 the ASTRAL analyses (RF 36). The addition of paralogs reduced disagreement in both cases
423 (supermatrix RF 38; ASTRAL RF 26).

424 *Analysis*— The greatest differences among the 12 trees were between ASTRAL trees and
425 the supermatrix trees (Figures 4, S6, Table S3), with a mean RF distance between the six
426 supermatrix trees and the six ASTRAL trees of 78. Again, the addition of additional sequences in
427 the form of noncoding regions or paralogs reduced the disagreement between supermatrix and
428 ASTRAL analyses; the average RF distance for exons/noparalogs was 85, exons+paralogs 77,
429 supercontig/noparalogs 71, and supercontigs+paralogs 70. Agreement was higher among the
430 ASTRAL trees (mean RF 21, 138/159 nodes in agreement) than among the supermatrix trees
431 (mean RF 48, 116/159 nodes in agreement). The differences (RF 66) between ASTRAL and

supermatrix analyses for the full dataset (supercontigs+paralogs for all genes) at the species level can be ascribed to mostly minor repositionings of two outgroup taxa (*Bagassa guianensis* Aubl. and *Batocarpus orinoceros* H. Karst.) and 14 ingroup taxa (*A. hypargyreus* Hance, *A. thailandicus* C.C. Berg., *A. gomezianus ssp. gomezianus, A. sp. aff. lacucha, A. fulvicortex* F.M. Jarrett, *A. sp. aff nitidus, A. sp. aff fretesii, A. ovatus, A. rubrovenius* Warb., *A. pinnatisectus, A. cf. horridus, A. cf. camansi, A. excelsus* F.M. Jarrett, *A. jarrettiae* Kochummen) (Figure 6).

Phylogenetic resolution

Bootstrap support was high for all main analysis trees. Percent resolution was 90–95% for all supermatrix trees in the main analysis and did not differ materially between analyses. Among ASTRAL trees, percent resolution was between 84% and 97% for all analyses. By slight margins, the best-resolved trees for both supermatrix and ASTRAL analyses were those based on the largest dataset (Figure 6). Percent resolution based on quartet support for ASTRAL trees was between 57% and 60%. Significant conflict existed at the gene tree level (Table S4, Figure 6). For percent resolution measured by gene tree support (percentage of nodes supported by at least half of the 517 gene trees), scores ranged from 17–24%. In general, analyses including paralogs had reduced gene tree support, and the trees based on supercontigs with no paralogs had the highest scores (24% for both ASTRAL and supermatrix).

A more detailed analysis of the differences between the *exon* and *supercontig* datasets revealed that even if the final species trees had similar resolution, the *supercontig* trees were based on more information because the gene trees were significantly more informative. A non-parametric Wilcoxon test indicated that the inclusion of non-coding sequences significantly increased both the mean bootstrap support (+5.47, $P < 0.0001$) and the number of splits with

455 over 30% support (+8, $P = 0.0022$). Because nodes under 30% were collapsed for species tree
456 estimation, the species tree in the *supercontig* dataset was based on 9% more splits across the
457 517 gene trees (total of 51,307) than the species tree in the *exon* dataset (total 47,067). These
458 patterns persisted in the no-paralog datasets (difference in mean bootstrap support: +2.06, $P <$
459 0.0001; difference in nodes over 30%: +9, $P < 0.0001$; overall difference in splits for 333
460 collapsed trees: 36,373 vs. 33,404 or 9%). Because the addition of non-coding sequences also
461 increased agreement between supermatrix and ASTRAL analyses (see above), this suggests that
462 at least some disagreement between supermatrix and species-tree analyses arises not only from
463 incomplete lineage sorting but also from lack of resolution at the gene tree level, something that
464 has also been observed at deeper phylogenetic scales (Pease et al. 2018).
465
466 *Phylogenetic relationships*
467 The genus *Artocarpus* was monophyletic in all 12 main analyses, as were subgenera
468 *Cauliflori* F.M. Jarrett and *Prainea* (King) Zerega, Supardi, & Motley (Table 1). Subgenus
469 *Artocarpus* was monophyletic excluding *A. sepicanus* Diels, and subgenus *Pseudojaca* was
470 monophyletic excluding *A. altissimus* J.J. Smith. In all supermatrix analyses and five ASTRAL
471 analyses, *Artocarpus sepicanus* Diels and *A. altissimus* J.J. Smith formed a clade sister to
472 subgenera *Cauliflori* and *Artocarpus*; however, in the codon-partitioned ASTRAL analyses, they
473 formed a grade in the same position, and in the ASTRAL supercontig analysis, only *A. altissimus*
474 was in that position, while *A. sepicanus* was sister to subgenus *Pseudojaca*. The backbone
475 phylogeny was otherwise identical in all twelve trees: subgenus *Prainea* was sister to all other
476 *Artocarpus*, which comprised a grade in this order: subgenus *Pseudojaca*, *A. sepicanus* + *A.*
477 *altissimus* (usually), followed by subgenus *Cauliflori* + subgenus *Artocarpus*. Apart from the

478 monophyly of the genus, which was supported by 61% of gene trees in the complete dataset
479 (supercontig, all genes), subgeneric relationships had much less support at the gene tree level.
480 The position of subg. *Prainea* was supported by 28% of gene trees; subg. *Pseudojaca* by 7%,
481 and subgenera *Artocarpus/Cauliflori* by only 4%. Quartet support, however, was substantially
482 higher (Figure 6).

483 Within subgenus *Artocarpus*, both of Jarrett's sections were monphyletic (leaving aside
484 *A. sepicanus, A, hirsutus*, and *A. nobilis*, which she considered anomalous and did not place in
485 sections), but none of the five series were monophyletic. The closest was series *Rugosi* F.M.
486 Jarrett which is characterized by rugose (sulcate to tuberculate) staminate inflorescences. It was
487 nearly monophyletic in most analyses, requiring only the exclusion of one rugose species (*A.*
488 *obtusus* F.M. Jarrett) and the inclusion of one non-rugose species (*A. teijsmannii* Miq.). Members
489 of series *Incisifolii* F.M. Jarrett, characterized by incised adult leaves (as with breadfruit, *A.*
490 *altilis*), formed two non-sister monophyletic clades, one in the Philippines and one ranging from
491 Indonesia to Oceania, both containing several potential undescribed species. Within subgenus
492 *Pseudojaca*, section *Pseudojaca* was monophyletic (excluding *A. altissimus*), as was series
493 *Clavati* F.M. Jarrett—characterized by clavate interfloral bracts. Series *Peltati* F.M. Jarrett—
494 characterized by peltate interfloral bracts—would be monophyletic if *A. tonkinensis* A. Chev.
495 were excluded, the latter species being sister to series *Clavati* in all main analyses.

496 Most species (for which we included at least two samples) were monophyletic as well,
497 but several were not monophyletic in any analysis, including *A. treculianus* Elmer, *A.*
498 *sarawakensis* F.M. Jarrett, *A. lanceifolius* Roxb., *A. rigidus* Blume, *A. teijsmanni*, and *A. nitidus*
499 Tréc. The type of *A. teijsmannii* ssp. *subglabrus* C.C. Berg was sister to *A. sepicanus* in all

analyses, while *ssp. teijsmannii* and other accessions of *ssp. subglabrus* were elsewhere within subgenus *Artocarpus*.

The neotropical Artocarpeae formed a clade sister to *Artocarpus* in all twelve trees. While *Batocarpus* was monophyletic in all supermatrix analyses, usually nested within a grade comprising *Clarisia*, neither *Batocarpus* nor *Clarisia* was monophyletic in any ASTRAL tree.

DISCUSSION

Taxon sampling

Although other studies have successfully applied target enrichment to recover sequences from herbarium and museum material (Guschanski et al. 2013; Hart et al. 2016), to our knowledge, this is among the first to use herbarium collections to achieve near-complete taxon sampling in a tropical plant genus of this size (ca. 70 spp.). The ability to successfully sequence herbarium material was indispensable for this study. For 34 of 90 (38%) ingroup taxa remaining in the final analyses (including subspecies and the nine individuals of uncertain affinities), we did not have access to any fresh or silica-dried material and relied exclusively on herbarium specimens. In some cases, the only readily available samples were approximately 100 years old, as in the case of *A. treculianus* sensu stricto (coll. 1910–1911: 369–370 genes recovered after filtering), *A. nigrescens* Elmer (coll. 1919: 431 genes), and *A. pinnatisectus* (type coll. 1913: 425 genes). Although old samples certainly had a lower success rate than silica-dried material, and sample degradation undoubtedly contributed to shorter assembled contigs, age alone was not significantly associated with recovery of fewer loci (Fig. 3). We hope these results serve as

encouragement for others to aim for complete taxon sampling with minimally-destructive sampling from natural history collections when newly-collected material is not available.

While fieldwork remains among the most important aspects for systematic biology studies, phylogenetic reconstruction can benefit dramatically from the incorporation of DNA from museum specimens. In this study, we were able to sequence several DNA extractions that had been prepared several years ago for previous studies but had been unusable, because PCR amplification for Sanger sequencing failed (presumably due to small fragment size) (Zerega et al. 2010; Williams et al., 2017). The ability to achieve near-complete taxon sampling from museum material will open up new opportunities for phylogeny-based analyses of clades with species that are difficult to collect, rare, or extinct, but present in herbarium collections. Our results suggest that near-complete taxon sampling can improve consistency between analyses, resulting in more reliable phylogenies. A previous study (Kates et al. 2018) using a smaller dataset of 22 *Artocarpus* species, found substantial disagreement between analyses in the backbone phylogeny of *Artocarpus*, in particular the positions of *Prainea* and *A. sepicanus.* Here, all 12 main analyses recovered almost the same backbone, disagreeing occasionally as to the positions of *A. altissimus* and *A. sepicanus*. Others have likewise found that missing taxa can substantially impact phylogenetic reconstructions (de la Torre-Bárcena et al. 2009). Robust taxon sampling also has serious implications for biodiversity conservation. *Artocarpus treculianus* is listed as Vulnerable by the IUCN (World Conservation Monitoring Centre 1998). Due to the availability of sequences from century-old herbarium sheets, we now know that this species is not monophyletic and that the two obsolete taxa (*A. ovatifolius* Merr. and *A. nigrescens*, an unusual taxon with black fruits) that Jarrett sunk into *A. treculianus* (Jarrett 1959) should probably be reinstated. Splitting a Vulnerable species into three will result, at the very least, in three Vulnerable species. The

545 availability of material from collections has also revealed new species including *A. bergii* E.M.
546 Gardner, Zerega, and Arifiani (Gardner et al., in review), a close ally of breadfruit from the
547 Maluku Islands and *A. vietnamicus* E.M. Gardner and N.J.C. Zerega (Gardner and Zerega, in
548 review), a montane species endemic to Vietnam that resembles *A. excelsus* F.M. Jarrett and *A.*
549 *lowii* King (*A. aff. excelsus*).
550
551 *Impact of various analysis methods*
552 Of all the analytical variants we tested, partitioning the "exon" analyses by codon
553 position had the least impact, resulting in no major topological changes in any analysis. This is
554 not surprising, given that RAxML's GTRCAT model provides for rate heterogeneity even absent
555 explicit partitioning (Stamatakis 2006). Comparisons of analyses with and without paralogs, with
556 and without non-coding sequences, and ASTRAL versus supermatrix revealed moderate
557 disagreement, mostly at shallow phylogenetic depths. However, in all cases, disagreement
558 decreased if additional sequences were added to a dataset (i.e., adding paralogs to the non-coding
559 present/absent comparison, adding non-coding sequences to the paralogs/no-paralogs
560 comparison, or adding either non-coding sequences or paralogs to the supermatrix/ASTRAL
561 comparison). This suggests that more data can lead to a certain amount of convergence in
562 analyses, even though simply adding more data to a supermatrix may not improve the accuracy
563 of the resulting species tree (Degnan and Rosenberg 2009).
564 Although adding or extending loci may reduce disagreement between analyses, it may
565 not always increase phylogenetic resolution. Certainly, with small numbers of genes, there may
566 not be enough informative characters to resolve a phylogeny, and resolution may increase as loci
567 are added. With hundreds of genes, however, lack of informative characters is not the problem.

568 Here, although adding paralogs slightly reduced disagreement between analyses, overall percent
569 resolution remained static, and gene tree concordance with the species tree decreased. Other
570 phylogenomic studies have also found high rates of gene tree discordance (Degnan and
571 Rosenberg 2009; Wickett et al. 2014; Copetti et al. 2017; Pease et al. 2018; Liu et al. 2019).
572 Gene tree discordance results from, inter alia, biological processes such as incomplete lineage
573 sorting or ancient hybridization, and therefore may reflect not a lack of phylogenetic resolution,
574 but the non-existence of a fixed, absolute species tree. Nonetheless, just as bootstrap support can
575 convey a misleading sense of certainty, support measured by the rate of gene-tree support can
576 exaggerate uncertainty. For example, if a gene tree generally supports a clade, but has one out-
577 of-place taxon, perhaps due to an incomplete or erroneous sequence, that gene tree will not be
578 counted as supporting the clade in question. Support measure as the proportion of gene tree
579 quartets supporting each node, not the frequency of the exact clade being tested, may provide a
580 more realistic measure of support (Sayyari and Mirarab 2016); in our analyses, they were
581 generally lower than bootstrap values but substantially higher than gene-tree support.
582 Based on these results, we conclude that partitioning by codon position is not necessary
583 for our data, and analyses at similar phylogenetic scales may also not benefit from such a
584 partitioning scheme, at least for analyses that provide for built-in rate heterogeneity such as
585 RAxML. We do however recommend that when possible, flanking non-coding sequences be
586 included in analyses. The benefits of more informative gene trees, and thus more reliable final
587 analyses, outweigh any minimal advantage gained in partitioning by codon position, at least for a
588 data set like ours. Finally, in light of the increased congruence between analyses as our data set
589 was enlarged, we suggest using as many loci and as much flanking noncoding sequence as is
590 available, with the caveat to exercise caution with regard to taxa with excessive missing data.

The cutoffs we used, >20% of the average sequence length and >~20% of loci, might be made more stringent, as some inter-analysis disagreement appeared to center around samples with more missing data.

Taxonomic considerations

Our results provide a phylogenetic framework for a taxonomic revision of *Artocarpus*, currently in progress. The subgeneric divisions made by Jarrett (1959, 1960) and Zerega et al. (2010) can be maintained with minor modifications to account for the anomalous *A. sepicanus* and *A. altissimus*, which in 9/12 main analyses formed a clade. It is extremely curious that these two species should be closely allied, and the occasional disagreement as to their affinity in the ASTRAL analyses warrants further investigation. *Artocarpus sepicanus* is a fairly ordinary member of subgenus *Artocarpus*, remarkable only in that its characters seem intermediate between those defining section *Artocarpus* and section *Duricarpus* F.M. Jarrett (Jarrett 1959). *Artocarpus altissimus*, on the other hand, is a most unusual member of subgenus *Pseudojaca*, placed in that division solely on the basis of distichous leaves, and the fusion of perianth tissue of adjacent carpellate flowers. Its leaves, with their trinerved bases and glandular-crenate margins, are unique in the genus *Artocarpus* and are reminiscent of *Morus*, reflected in its basionym *Morus altissima* Miq. Anatomical studies may reveal more, but the only apparent morphological affinity between *A. sepicanus* and *A. altissimus* are bifid styles, a moraceous plesiomorphy (Clement and Weiblen 2009) present occasionally in subgenus *Artocarpus* but unique to *A. altissimus* in subgenus *Pseudojaca*. The leaf margins of *A. sepicanus*, which are entire in mature trees but toothed in juveniles, are perhaps reminiscent of the glandular-crenate leaves of *A. altissimus*.

In addition, the phylogeny supports the broad outlines of Jarrett's (1959, 1960) sections, validating her careful morphological and anatomical studies, which built on those of Renner (1907). The sections within subgenera *Artocarpus* might be maintained with the minor adjustment of including *A. hirsutus* Lam. and *A. nobilis* Thwaites in section *Duricarpus*. Jarrett noted that those species had characters intermediate between sections *Artoarpus* and *Duriarpus*, and indeed, their positions in all main analyses was that of sister to most of the rest of section *Duricarpus*, possibly indicating the preservation of plesiomorphic characters. The two subspecies of *A. lanceifolius* Roxb. were not sister taxa and may be distinguished by varying stamen lengths and leaf sizes. In *A. sarawakensis*, the Sumatran individual was not sister to the sample from Sarawak and the former might more appropriately be considered a variant of *A. lancefolius ssp. lanceifolius*, to which it bears similarity in syncarp characters, leaf shape, differing only in the presence of a dense indumentum on the stipules.

At the series level within subgenus *Artocarpus*, a wholesale reconsideration is probably necessary, although in a rough sense, the relevant characters match the clades. The exception is series *Angusticarpi*, which did not form a consistent clade or grade. *Artoarpus teijsmannii* belongs with the species of series *Rugosi*, which clade, in the broad sense, is characterized by species with either rugose staminate inflorescences and/or dimorphic perianths (long+short) on carpellate inflorescences, the latter of which applies to *A. teijsmannii* as well. The remaining species, *A. lowii*, has morphological affinities to *A. excelsus* and *A. vietnamicus*. All three species have smallish syncarps (to 6.5–7 cm), with shallow surface protrusions (except in *A. vietnamicus*), and smallish elliptical leaves (6–36 cm long). These characters are closer to those of *A. sepicanus* than most other members of section *Artocarpus*, which often have broadly ovate to obovate leaves (up to ca. 70 cm long), suggesting that the three species, scattered throughout

637 section *Artocarpus*, may preserve some plesiomorphic characters. The two non-sister clades
638 comprising series *Incisifolii* do indeed contain all species with incised adult leaves, but many
639 others as well; they are defined more by geography than by morphology. The clade containing *A.*
640 *altilis* contains perhaps three new species, of interest as previously-unknown wild relatives of
641 breadfruit. These include *A. bergii,* known only from the Maluku Islands, and two accessions of
642 uncertain affinity from the living collections of Bogor Botanical Gardens (*cf. camansi* and *cf.*
643 *horridus*) that formed a clade yet have starkly divergent vegetative morphology, requiring further
644 study to determine whether they are distinct taxa. The status of *A. horridus*— morphologically
645 similar to *A. camansi*—is unclear; one accession fell in its expected place (sister to *A. camansi*),
646 but the position of the other, sister to the entire clade, must be treated with caution, as that
647 sample had among the highest proportions of missing data.

648 Within subgenus *Pseudojaca*, to the extent we included multiple accessions per species,
649 our results mostly supported Jarrett's (1960) revision. The series were largely monophyletic,
650 with the exception of the position of *A. tonkinensis* (with peltate interfloral bracts) nested within
651 the clade distinguished by clavate interfloral bracts. The ancestral state for interfloral bracts is
652 likely peltate (Clement and Weiblen 2009), so *A. tonkinensis* may simply represent a
653 plesiomorphic taxon sister to a derived clade. As Williams et al. (2017) found, the four species
654 sunk into *A. lacucha* by Berg et al. (2006) (*A. dadah* Miq., *A. ovatus* Blanco, *A. fretesii,* and *A.*
655 *vrieseanus* var. *refractus*) do not belong together. The varieties of *A. vrieseanus* do in fact form a
656 clade, and *A. longifolius* ssp. *adpressus* C.C. Berg, described by Berg (2005) ahead of the *Flora*
657 *Malesiana*, does indeed belong with the type subspecies. However, *A. gomezianus* ssp.
658 *zeylanicus* Jarrett was not sister to the type subspecies, but instead apparently belonged with *A.*
659 *lacucha* sensu Jarrett. *Artocarpus nitidus* ssp. *humilis* and ssp. *griffithii* usually formed a clade,

but otherwise the other two *A. nitidus* subspecies were unrelated. Among the Chinese species described since Jarrett's (1960) revision, the rather distinctive *A. pithecogallus* C.Y. Wu and *A. gongshanensis* S. K. Wu ex C. Y. Wu & S. S. Chang fell in the expected clavate-bracted clade. Our sampling did not include *A. nanchuanensis*, but this species is morphologically similar to *A. hypargyreus*, and subsequent sequencing after the main analyses were complete confirmed the affinity. We were unable to successfully sequence *A. nigrifolius*, but an examination of type specimen confirmed a close affinity to *A. hypargyreus*. Although the original description mentions possible affinities with *A. styracifolius*, the clavate rather than ovoid staminate inflorescences are much closer to *A. hypargyreus*. Moreover, the primary distinguishing characters—leaves drying black—is frequently found on *A. hypargyreus* herbarium sheets as well, including the type of the latter species (*Hance 4484*, P) and might therefore not be a proper diagnostic.

CONCLUSION

The increasing availability of phylogenomic datasets has dramatically changed the practice of revisionary systematics. Data sets containing hundreds or thousands of loci produce trees with extremely high statistical support, apparently providing ironclad frameworks for making taxonomic decisions. However, apparent high support for relationships may often be an artifact of the massive number of characters available for phylogenetic inference, masking real uncertainties that are revealed only by employing a variety of analytical methods. By the same token, focusing on exclusively conserved coding regions—an inherent feature of some reference-based assembly methods—can result in unnecessarily uninformative gene trees, leading to poor support at the species tree level. Using a data set with near-complete taxon sampling, we

683 demonstrated that decisions made in how to conduct analyses can substantially affect
684 phylogenetic reconstruction, resulting in discordant phylogenies, each with high statistical
685 support. Employing multiple analytical methods can help separate truly robust phylogenetic
686 relationships from those that only appear to be well-supported but are not consistent across
687 analyses. While codon partitioning and model choice did not substantially alter our phylogeny,
688 the inclusion of flanking non-coding sequences in analyses significantly increased the number of
689 informative splits at the gene tree level, resulting ultimately in more robust species trees. In
690 general, increasing the size of datasets through the inclusion of paralogous genes increased
691 convergence between analysis methods but did not reduce gene tree conflict, which likely
692 resulted from biological and not analytical processes; for this reason, we prefer quartet-based
693 scoring methods as the most informative ways of determining support for species trees.
694
695 We provide a robust phylogenetic framework for *Artocarpus*, making use of herbarium
696 specimens up to 106 years old to supplement our own collections and achieve near-complete
697 taxon sampling, demonstrating the value of even very old natural history collections in
698 improving phylogenetic studies. Our results will inform future evolutionary and systematic
699 studies of this important group of plants. More generally, the results may guide future analyses
700 of HybSeq datasets, particularly those combining fresh with museum material, by counseling
701 careful attention to dataset construction and analysis method to produce the most informative
702 phylogenetic hypotheses.
703
704 FUNDING

This work was supported by the United States National Science Foundation (DEB award numbers 0919119 to NJCZ, 1501373 to NJCZ and EMG, 1342873 and 1239992 to NJW, and DBI award number 1711391 to EMG); the Northwestern University Plant Biology and Conservation Program; The Initiative for Sustainability and Energy at Northwestern University; the Garden Club of America; the American Society of Plant Taxonomists; a Systematics Research Fund grant from the Linnaean Society and the Systematics Association; and the Botanical Society of America.

ACKNOWLEDGEMENTS

We thank Postar Miun, Jeisin Jumian, Markus Gubilil, Aloysius Laim, S. Brono, Jegong anak Suka, Salang anak Nyegang, Jugah anak Tagi, Wan Nuur Fatiha Wan Zakaria, and Harto for assistance in the field; the Pritzker Laboratory for Molecular Systematics at the Field Museum of Natural History (K. Feldheim) for the use of sequencing facilities; J. Fant, E. Williams, H. Noble, R. Overson, and B. Cooper for assistance in the lab; the Sabah Agriculture Department for access to collections; and the following herbaria for access to collections for examination and/or sampling: BM, BO, F, IBSC, K, FRIM, FTBG, L, MO, NY, PNH, SAN, SING, SAR, SNP.

LITERATURE CITED

van Bakel H., Stout J.M., Cote A.G., Tallon C.M., Sharpe A.G., Hughes T.R., Page J.E. 2011. The draft genome and transcriptome of *Cannabis sativa*. Genome Biol. 12:R102.

Bakker F.T., Lei D., Yu J., Mohammadin S., Wei Z., van de Kerke S., Gravendeel B.,

728 Nieuwenhuis M., Staats M., Alquezar-Planas D.E., Holmer R. 2016. Herbarium genomics:
729 Plastome sequence assembly from a range of herbarium specimens using an Iterative
730 Organelle Genome Assembly pipeline. Biol. J. Linn. Soc. 117:33–43.
731 Bankevich A., Nurk S., Antipov D., Gurevich A.A., Dvorkin M., Kulikov A.S., Lesin V.M.,
732 Nikolenko S.I., Pham S., Prjibelski A.D., Pyshkin A. V., Sirotkin A. V., Vyahhi N., Tesler
733 G., Alekseyev M.A., Pevzner P.A. 2012. SPAdes: A New Genome Assembly Algorithm
734 and Its Applications to Single-Cell Sequencing. J. Comput. Biol. 19:455–477.
735 Berg C. 2005. Flora Malesiana precursor for the treatment of Moraceae 8: other genera than
736 Ficus. Blumea. 50:535–550.
737 Brewer G.E., Clarkson J.J., Maurin O., Zuntini A.R., Barber V., Bellot S., Biggs N., Cowan R.S.,
738 Davies N.M.J., Dodsworth S., others. 2019. Factors affecting targeted sequencing of 353
739 nuclear genes from herbarium specimens spanning the diversity of angiosperms. Front.
740 Plant Sci. 10:1102.
741 Broad Institute. 2016. Picard tools. Available from
742 https://broadinstitute.github.io/picard/%5Cnhttp://broadinstitute.github.io/picard/.
743 Buerki S., Baker W.J. 2016. Collections-based research in the genomic era. Biol. J. Linn. Soc.
744 117:5–10.
745 Capella-Gutiérrez S., Silla-Martínez J.M., Gabaldón T. 2009. trimAl: A tool for automated
746 alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 25:1972–1973.
747 Castañeda-Álvarez N.P., Khoury C.K., Achicanoy H.A., Bernau V., Dempewolf H., Eastwood
748 R.J., Guarino L., Harker R.H., Jarvis A., Maxted N., Müller J. V., Ramirez-Villegas J., Sosa
749 C.C., Struik P.C., Vincent H., Toll J. 2016. Global conservation priorities for crop wild
750 relatives. Nat. Plants.:1–6.

751 Chamala S., García N., Godden G.T., Krishnakumar V., Jordon-Thaden I.E., De Smet R.,
752 Barbazuk W.B., Soltis D.E., Soltis P.S. 2015. MarkerMiner 1.0: A new application for
753 phylogenetic marker development using angiosperm transcriptomes. Appl. Plant Sci.
754 3:1400115.
755 Clement W.L., Weiblen G.D. 2009. Morphological Evolution in the Mulberry Family
756 (Moraceae). Syst. Bot. 34:530–552.
757 Copetti D., Búrquez A., Bustamante E., Charboneau J.L.M., Childs K.L., Eguiarte L.E., Lee S.,
758 Liu T.L., McMahon M.M., Whiteman N.K., Wing R.A., Wojciechowski M.F., Sanderson
759 M.J. 2017. Extensive gene tree discordance and hemiplasy shaped the genomes of North
760 American columnar cacti. Proc. Natl. Acad. Sci. U. S. A.
761 Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the
762 multispecies coalescent. Trends Ecol. Evol.
763 Doyle J., Doyle J. 1987. Genomic plant DNA preparation from fresh tissue-CTAB method.
764 Phytochem. Bull. 19:11–15.
765 Faircloth B.C. 2015. PHYLUCE is a software package for the analysis of conserved genomic
766 loci. Bioinformatics. 32:786–788.
767 Faircloth B.C., McCormack J.E., Crawford N.G., Harvey M.G., Brumfield R.T., Glenn T.C.
768 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple
769 evolutionary timescales. Syst. Biol. 61:717–26.
770 Gardner E.M., Arifiani D., Zerega N.J.C. In review. *Artocarpus bergii* (Moraceae): a new
771 species in the breadfruit clade from the Moluccas. Syst. Bot.
772 Gardner E.M., Johnson M.G., Ragone D., Wickett N.J., Zerega N.J.C. 2016. Low-Coverage,
773 Whole-Genome Sequencing of *Artocarpus camansi* (Moraceae) for Phylogenetic Marker

774 Development and Gene Discovery. Appl. Plant Sci. 4:1600017.

775 Gardner E.M., Zerega N.J.C. In review. *Artocarpus vietnamicus* (Moraceae): a new species from
776 the mountain forests of Vietnam. Phytotaxa.

777 Guschanski K., Krause J., Sawyer S., Valente L.M., Bailey S., Finstermeier K., Sabin R.,
778 Gilissen E., Sonet G., Nagy Z.T., Lenglet G., Mayer F., Savolainen V. 2013. Next-
779 generation museomics disentangles one of the largest primate radiations. Syst. Biol.
780 62:539–554.

781 Hart M.L., Forrest L.L., Nicholls J.A., Kidner C.A. 2016. Retrieval of hundreds of nuclear loci
782 from herbarium specimens. Taxon. 65:1081–1092.

783 He N., Zhang C., Qi X., Zhao S., Tao Y. 2013. Draft genome sequence of the mulberry tree
784 *Morus notabilis*. Nat. Commun. 4:2445.

785 Jarrett F.M. 1975. Four new *Artocarpus* species from Indo-Malesia (Moraceae). Blumea.
786 22:409–410.

787 Johnson M.G., Gardner E.M., Liu Y., Medina R., Goffinet B., Shaw A.J., Zerega N.J.C., Wickett
788 N.J. 2016. HybPiper: Extracting Coding Sequence and Introns for Phylogenetics from High-
789 Throughput Sequencing Reads Using Target Enrichment. Appl. Plant Sci. 4:1600016.

790 Johnson M.G., Pokorny L., Dodsworth S., Botigué L.R., Cowan R.S., Devault A., Eiserhardt
791 W.L., Epitawalage N., Forest F., Kim J.T., Leebens-Mack J.H., Leitch I.J., Maurin O.,
792 Soltis D.E., Soltis P.S., Wong G.K., Baker W.J., Wickett N.J. 2019. A Universal Probe Set
793 for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using
794 k-Medoids Clustering. Syst. Biol. 68:594–606.

795 Kates H.R., Johnson M.G., Gardner E.M., Zerega N.J.C., Wickett N.J. 2018. Allele phasing has
796 minimal impact on phylogenetic reconstruction from targeted nuclear gene sequences in a

797 case study of *Artocarpus*. Am. J. Bot.

798 Katoh K., Standley D.M. 2013. MAFFT multiple sequence alignment software version 7:
799 improvements in performance and usability. Mol. Biol. Evol. 30:772–80.

800 Keller O., Kollmar M., Stanke M., Waack S. 2011. A novel hybrid gene prediction method
801 employing protein multiple sequence alignments. Bioinformatics. 27:757–63.

802 Kochummen K.M. 1998. New species and varieties of Moraceae from Malaysia. Gard. Bull.
803 Singapore. 50:197–219.

804 Kubatko L.S., Degnan J.H. 2007. Inconsistency of phylogenetic estimates from concatenated
805 data under coalescence. Syst. Biol. 56:17–24.

806 de la Torre-Bárcena J.E., Kolokotronis S.-O., Lee E.K., Stevenson D.W., Brenner E.D., Katari
807 M.S., Coruzzi G.M., DeSalle R. 2009. The impact of outgroup choice and missing data on
808 major seed plant phylogenetics using genome-wide EST data. PLoS One. 4:e5764.

809 Lanfear R., Calcott B., Kainer D., Mayer C., Stamatakis A. 2014. Selecting optimal partitioning
810 schemes for phylogenomic datasets. BMC Evol. Biol. 14:82.

811 Larsson A. 2014. AliView: A fast and lightweight alignment viewer and editor for large datasets.
812 Bioinformatics. 30:3276–3278.

813 Lee B.N. 2014. Solid Phase Reverse Immobilization (SPRI) Bead Technology for Micro RNA
814 Clean Up using the. .

815 Li H., Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform.
816 Bioinformatics. 25:1754–1760.

817 Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin
818 R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25:2078–
819 2079.

820 Liu Y., Johnson M.G., Cox C.J., Medina R., Devos N., Vanderpoorten A., Hedenäs L., Bell N.E.,
821 Shevock J.R., Aguero B., Quandt D., Wickett N.J., Shaw A.J., Goffinet B. 2019. Resolution
822 of the ordinal phylogeny of mosses using targeted exons from organellar and nuclear
823 genomes. Nat. Commun.
824 Mandel J.R., Dikow R.B., Funk V. a., Masalia R.R., Staton S.E., Kozik A., Michelmore R.W.,
825 Rieseberg L.H., Burke J.M. 2014. A Target Enrichment Method for Gathering Phylogenetic
826 Information from Hundreds of Loci: An Example from the Compositae. Appl. Plant Sci.
827 2:1300085.
828 Medina R., Johnson M.G., Liu Y., Wickett N.J., Shaw A.J., Goffinet B. 2019. Phylogenomic
829 delineation of *Physcomitrium* (Bryophyta: Funariaceae) based on targeted sequencing of
830 nuclear exons and their flanking regions rejects the retention of *Physcomitrella* ,
831 *Physcomitridium* and *Aphanorrhegma* . J. Syst. Evol.
832 Miller M.A., Pfeiffer W., Schwartz T. 2010. Creating the CIPRES Science Gateway for
833 inference of large phylogenetic trees. 2010 Gatew. Comput. Environ. Work. GCE 2010.
834 Mirarab S., Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many
835 hundreds of taxa and thousands of genes. Bioinformatics. 31:i44–i52.
836 Paradis E., Claude J., Strimmer K. 2004. APE: Analyses of phylogenetics and evolution in R
837 language. Bioinformatics. 20:289–290.
838 Pease J.B., Brown J.W., Walker J.F., Hinchliff C.E., Smith S.A. 2018. Quartet Sampling
839 distinguishes lack of support from conflicting support in the green plant tree of life. Am. J.
840 Bot.
841 Price M.N., Dehal P.S., Arkin A.P. 2009. Fasttree: Computing large minimum evolution trees
842 with profiles instead of a distance matrix. Mol. Biol. Evol. 26:1641–1650.

843 Prum R.O., Berv J.S., Dornburg A., Field D.J., Townsend J.P., Lemmon E.M., Lemmon A.R.
844 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA
845 sequencing. Nature. 526:569–573.
846 Quinlan A.R., Hall I.M. 2010. BEDTools: A flexible suite of utilities for comparing genomic
847 features. Bioinformatics. 26:841–842.
848 Rambaut A. 2016. FigTree v1.4.3. Mol. Evol. phylogenet ics Epidemiol.
849 Ranwez V. 2011. MACSE : Multiple Alignment of Coding SEquences Accounting for
850 Frameshifts and Stop Codons. 6.
851 Revell L.J. 2012. phytools: An R package for phylogenetic comparative biology (and other
852 things). Methods Ecol. Evol. 3:217–223.
853 Sarkar D. 2008. Lattice: Multivariate Data Visualization with R. New York: Springer.
854 Sayyari E., Mirarab S. 2016. Fast Coalescent-Based Computation of Local Branch Support from
855 Quartet Frequencies. Mol. Biol. Evol. 33:1654–1668.
856 Sayyari E., Whitfield J.B., Mirarab S. 2017. Fragmentary Gene Sequences Negatively Impact
857 Gene Tree and Species Tree Reconstruction. Mol. Biol. Evol. 34:3279–3291.
858 Schliep K.P. 2011. phangorn: Phylogenetic analysis in R. Bioinformatics. 27:592–593.
859 Smith S.A., Moore M.J., Brown J.W., Yang Y. 2015. Analysis of phylogenomic datasets reveals
860 conflict, concordance, and gene duplications with examples from animals and plants. BMC
861 Evol. Biol.
862 Staats M., Erkens R.H.J., van de Vossenberg B., Wieringa J.J., Kraaijeveld K., Stielow B., Geml
863 J., Richardson J.E., Bakker F.T. 2013. Genomic treasure troves: complete genome
864 sequencing of herbarium and insect museum specimens. PLoS One. 8:e69189.
865 Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with

866 thousands of taxa and mixed models. Bioinformatics. 22:2688–90.

867 Sukumaran J., Holder M.T. 2010. DendroPy: a Python library for phylogenetic computing.

868 Bioinformatics. 26:1569–71.

869 Tange O. 2018. GNU Parallel 2018. *https://doi.org/10.5281/zenodo.1146014*.

870 Team R.D.C. 2008. R: A language and environment for statistical computing. Vienna, Austria: R

871 Foundation for Statistical Computing.

872 Villaverde T., Pokorny L., Olsson S., Rincón-Barrado M., Johnson M.G., Gardner E.M., Wickett

873 N.J., Molero J., Riina R., Sanmartín I. 2018. Bridging the micro- and macroevolutionary

874 levels in phylogenomics: Hyb-Seq solves relationships from populations to species and

875 above. New Phytol.

876 Wang M.M.H., Gardner E.M., Chung R.C.K., Chew M.Y., Milan A.R., Pereira J.T., Zerega

877 N.J.C. 2018. Origin and diversity of an underutilized fruit tree crop, cempedak (Artocarpus

878 integer, Moraceae). Am. J. Bot.

879 Weitemier K., Straub S.C.K., Cronn R.C., Fishbein M., Schmickl R., McDonnell A., Liston A.

880 2014. Hyb-Seq: Combining Target Enrichment and Genome Skimming for Plant

881 Phylogenomics. Appl. Plant Sci. 2:1400042.

882 Wickett N.J., Mirarab S., Nguyen N., Warnow T., Carpenter E., Matasci N., Ayyampalayam S.,

883 Barker M.S., Burleigh J.G., Gitzendanner M.A., Ruhfel B.R., Wafula E., Der J.P., Graham

884 S.W., Mathews S., Melkonian M., Soltis D.E., Soltis P.S., Miles N.W., Rothfels C.J.,

885 Pokorny L., Shaw A.J., DeGironimo L., Stevenson D.W., Surek B., Villarreal J.C., Roure

886 B., Philippe H., DePamphilis C.W., Chen T., Deyholos M.K., Baucom R.S., Kutchan T.M.,

887 Augustin M.M., Wang J., Zhang Y., Tian Z., Yan Z., Wu X., Sun X., Wong G.K.-S.,

888 Leebens-Mack J. 2014. Phylotranscriptomic analysis of the origin and early diversification

889 of land plants. Proc. Natl. Acad. Sci. 111:E4859–E4868.

890 Williams E.W., Gardner E.M., Harris R., Chaveerach A., Pereira J.T., Zerega N.J.C. 2017. Out
891 of Borneo: Biogeography, phylogeny, and divergence date estimates of *Artocarpus*
892 (Moraceae). Ann. Bot. 119:611–627.

893 Witherup C., Zuberi M.I., Hossain S., Zerega N.J.C. 2019. Genetic Diversity of Bangladeshi
894 Jackfruit (*Artocarpus heterophyllus*) over Time and Across Seedling Sources. Econ. Bot.
895 73:233–248.

896 World Conservation Monitoring Centre. 1998. Artocarpus treculianus. Available from
897 http://dx.doi.org/10.2305/IUCN.UK.1998.RLTS.T33246A9771111.en.

898 Xi Z., Ruhfel B.R., Schaefer H., Amorim A.M., Sugumaran M., Wurdack K.J., Endress P.K.,
899 Matthews M.L., Stevens P.F., Mathews S., Davis C.C. 2012. Phylogenomics and a
900 posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. Proc.
901 Natl. Acad. Sci. U. S. A. 109:17519–24.

902 Zerega N.J.C., Nur Supardi M.N., Motley T.J. 2010. Phylogeny and Recircumscription of
903 Artocarpeae (Moraceae) with a Focus on *Artocarpus*. Syst. Bot. 35:766–782.

904 Zerega N.J.C., Wiesner-Hanks T., Ragone D., Irish B., Scheffler B., Simpson S., Zee F. 2015.
905 Diversity in the breadfruit complex (*Artocarpus*, Moraceae): genetic characterization of
906 critical germplasm. Tree Genet. Genomes. 11:1–26.

907 Zhang C., Sayyari E., Mirarab S. 2017. ASTRAL-III: Increased Scalability and Impacts of
908 Contracting Low Support Branches BT - Comparative Genomics: 15th International
909 Workshop, RECOMB CG 2017, Barcelona, Spain, October 4-6, 2017, Proceedings. In:
910 Meidanis J., Nakhleh L., editors. Cham: Springer International Publishing. p. 53–75.

911 Zhengyi W., Xiushi Z. 1989. TAXA NOVA NONNULLA MORACEARUM SINENSIUM.

912 Acta Bot. Yunnanica. 11:24–34.

913

914
915

Table 1. A summary of *Artocarpus* taxonomy following Zerega et al. (2010) at the subgeneric level, and Jarrett (1959–1960) at the section and series level. Species marked with an asterisk (*) were described after Jarrett's revision; we have listed them with the taxonomic divisions in which they place based on the phylogeny presented in this study. Species in gray text were not included in the phylogeny.

Subgenus	Section	Series	Species	Monophyletic
Artocarpus				Yes, if *A. sepicanus* is excluded
	Artocarpus			Yes, if *A. sepicanus* is excluded
		Angusticarpi	*A. lowii, A. teijsmannii*	Yes, if *A. teijsmanii* ssp. *subglabrous* is excluded, but see also series *Rugosi* for comments
		Incisifolii	[*A. altilis, A. camansi, A. mariannensis, A. horridus, A. bergii**], [*A. blancoi, A. treculianus, A. pinnatisectus, A. multifidus*]	No, but it consists of two monophyletic clades (separated by brackets to the left) defined by geography
		Rugosi	*A. scortechinii, A. elasticus, A. sericicarpus, A. tamaran, A. sumatranus, A. kemando, A. maingayi, A. corneri**, *A. jarrettiae**, *A. excelsus**, *A. obtusus**	In most analyses, yes if *A. lowii*, and *A. teijsmannii* ssp. *teijsmannii* are included
		Unplaced	*A. hirsutus, A. nobilis, A. sepicanus, A. vietnamicus**	
	Duricarpus			Yes, if *A. hirsutus* and *A. nobilis* are included
		Asperifolii	*A. melinoxylus, A. odoratissimus, A. hispidus, A. rigidus, A. chama, A. brevidpedunculatus, A. sarawakensis**	Yes, if *A. hirsutus* and *A. nobilis* are included, and *A. sarawakensis* and *A. brevipedunculatus* are excluded
		Laevifolii	*A. anisophyllus, A. lanceifolius*	Yes, if *A. sarawakensis* and *A. brevipedunculatus* are included
Cauliflori			*A. heterophyllus, A. integer, A. annulatus**	Yes
Pseudojaca				Yes, if *A. altissimus* is excluded.
	Glandulifolium		*A. altissimus*	Yes
	Pseudojaca	*Clavati*	*A. hypargyraeus, A. styracifolius, A. petelotii, A. pithecogallus**, *A. gongshanensis**, *A. nigrifolius**, *A. nanchuanensis**	Yes, if *A. tonkinensis* is included.
		Peltati	*A. glaucus, A. vrieseanus, A. xanthocarpus, A. longifolius, A. subrotundifolius, A. reticulatus, A. lacucha, A. gomezianus, A. tomentosulus, A. ovatus, A. tonkinensis, A. fretessii, A. dadah, A. rubrovenius, A. nitidus, A. fulvicortex, A. lacucha, A. albobrunneus**, *A. thailandicus**, *A. primackii**	Yes, if *A. tonkinensis* is excluded
Prainea			*A. limpato, A. papuanus, A. scandens, A. frutescens*	Yes

FIGURES

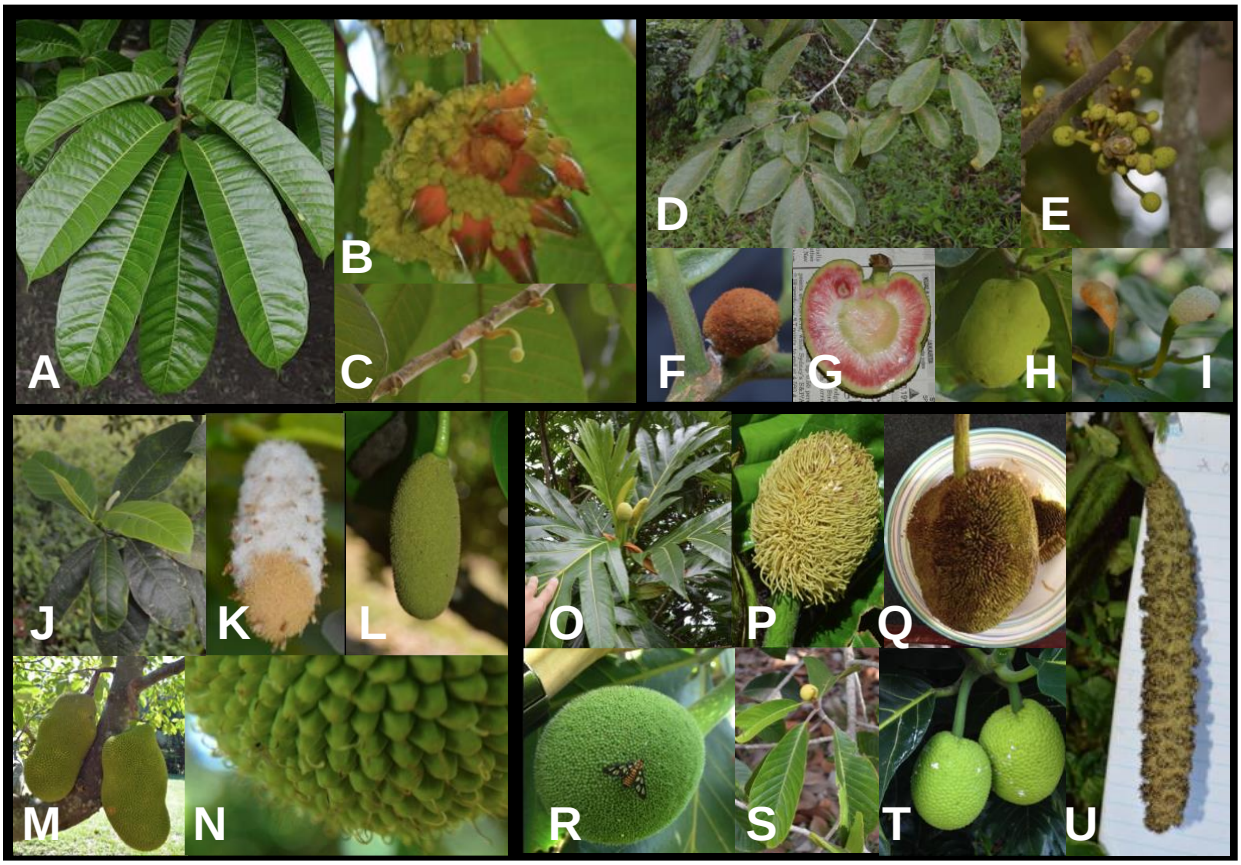


Figure 1. Diversity of *Artocarpus*. **Subg.** *Prainea*— (A) leaves, (B) syncarp, and (C) immature inflorescences of *A. limpato.* **Subg.** *Pseudojaca* — (D) leaves and (E) staminate inflorescences of *A. fretessii;* (F) carpellate inflorescence of *A. nitidus ssp. borneensis;* (G) syncarp of *A. primackii;* (H) syncarp of *A. nitidus ssp. lingnanensis;* and (I) staminate (left) and carpellate (right) inflorescences of *A. hypargyreus.* **Subg.** *Cauliflori*— (J) leaves of *A. integer*; (K–L) staminate inflorescences, (M) syncarps, and (N) carpellate inflorescence of *A. heterophyllus*. **Subg.** *Artocarpus*— (O) leaves and inflorescences of *A. altilis*; (P) carpellate inflorescence of *A. tamaran;* (Q) syncarp and (R) carpellate inflorescence of *A. odoratissimus*; (S) leaves and staminate inflorescence of *A. rigidus*; (T) syncarps of *A. altilis*; and (U) staminate inflorescence of *A. tamaran*.

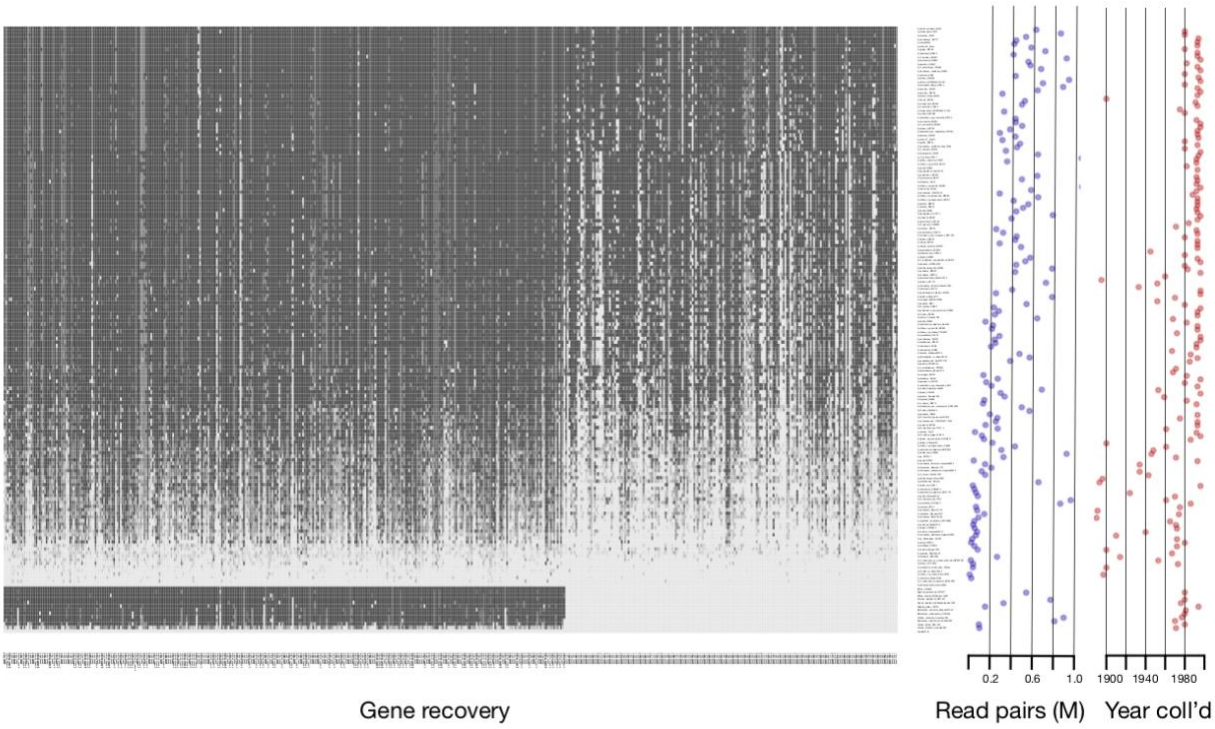
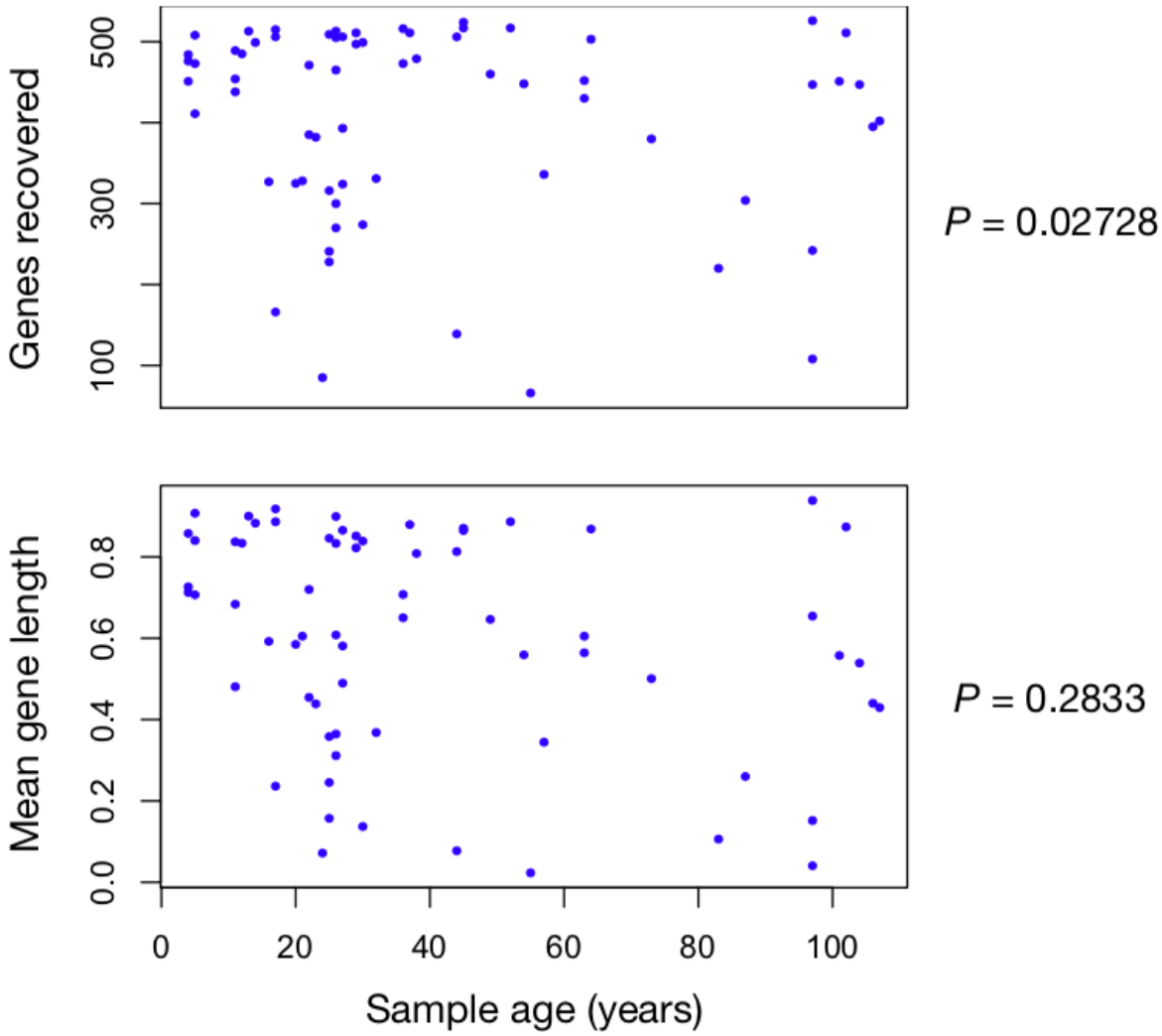


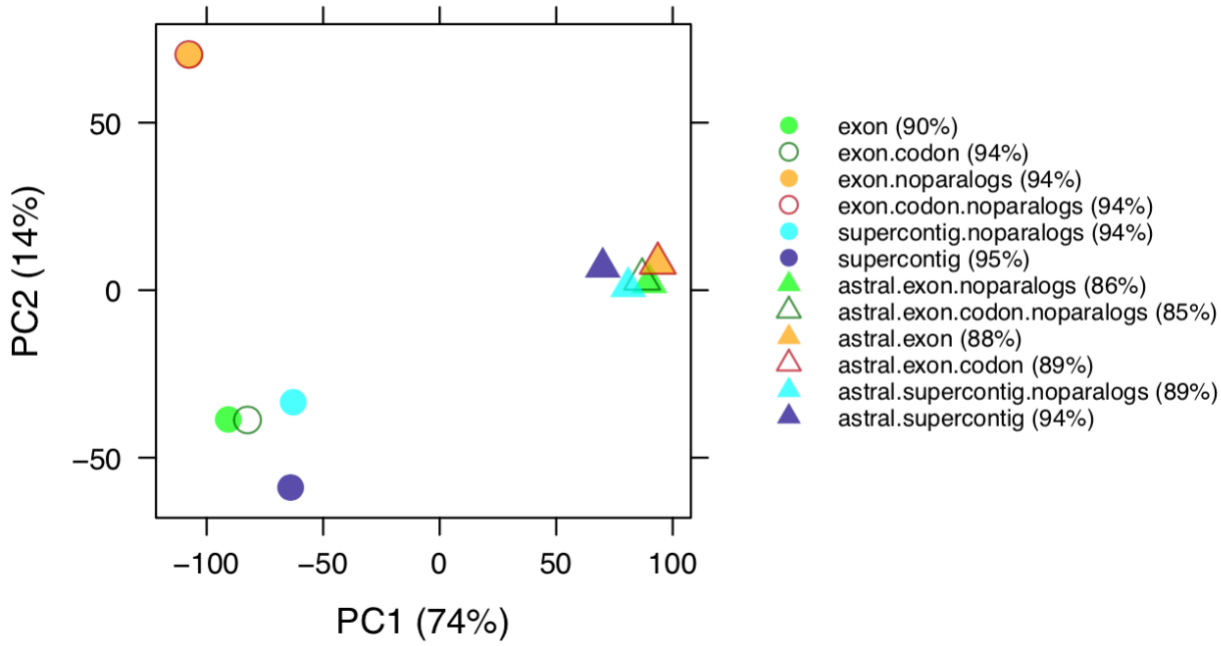
Figure 2. (A) Heatmap of gene recovery as a percentage of the average recovered sequence length for each gene. Rows represent samples, and columns represent genes. Darker colors indicate more complete recovery; white indicates no recovery. (B) Age (x-axis) for each sample. (C) Number of reads on target (x-axis) for each sample.



935

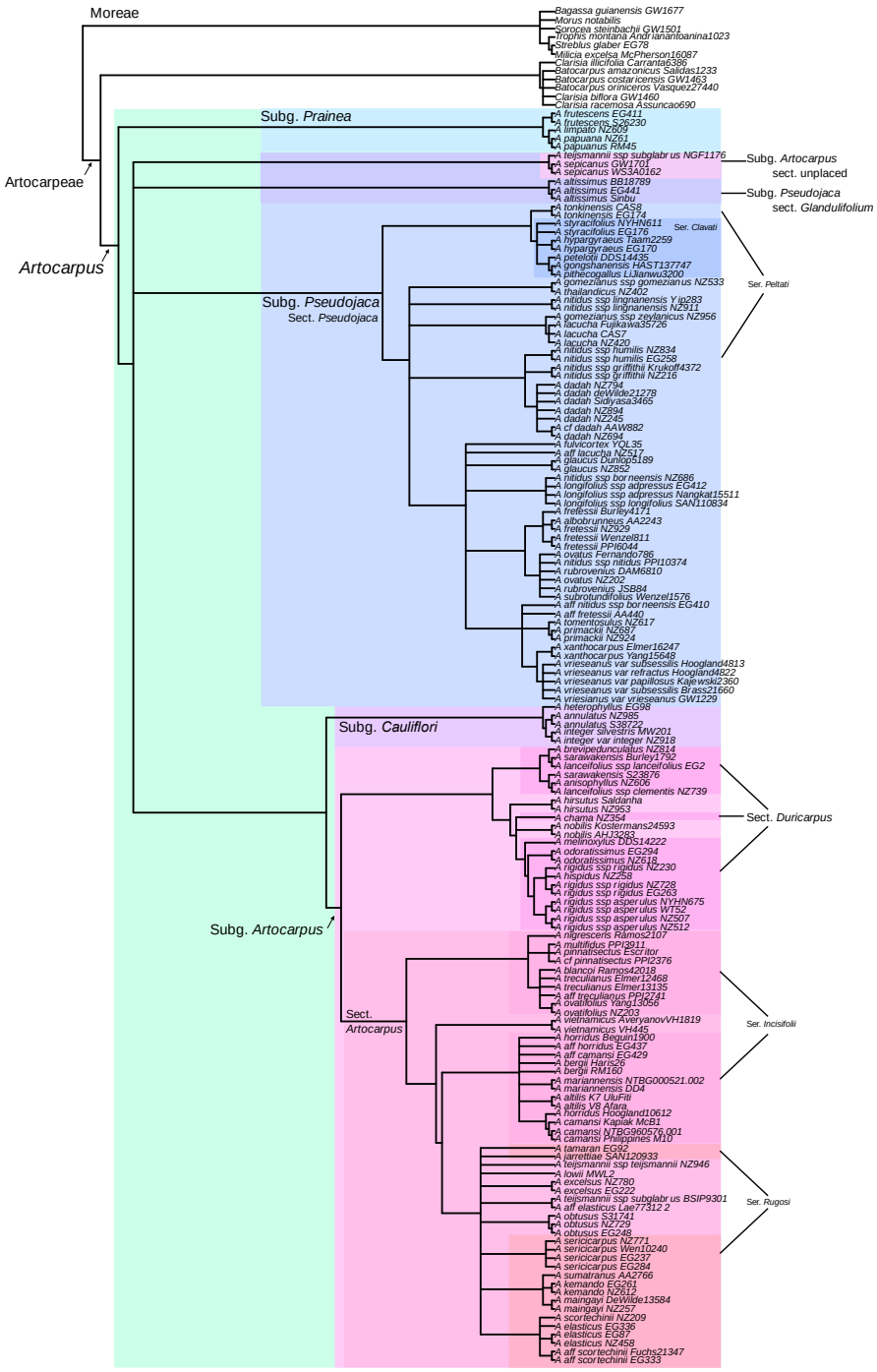
936 Figure 3. Comparison between herbarium sample age and total genes recovered (top) and mean

937 gene length as a proportion of average gene length for each gene (bottom)



938

939 Figure 4. PCA of Robinson-Foulds (RF) distances between all 12 main analyses.



940

941 Figure 5. Strict consensus of all 12 main-analysis trees (excluding only those analyses in which

942 exons and introns were aligned separately, for reasons discussed in the text). Colored boxes
943 reflect Jarrett's (1959, 1960) taxonomic divisions, as modified by Zerega et al. (2010).
944 Recently-described taxa that were split from older taxa recognized by Jarrett are classified
945 according to Jarrett's species concepts. Labels to the right of the tree denote major non-
946 monophyletic taxonomic divisions
947

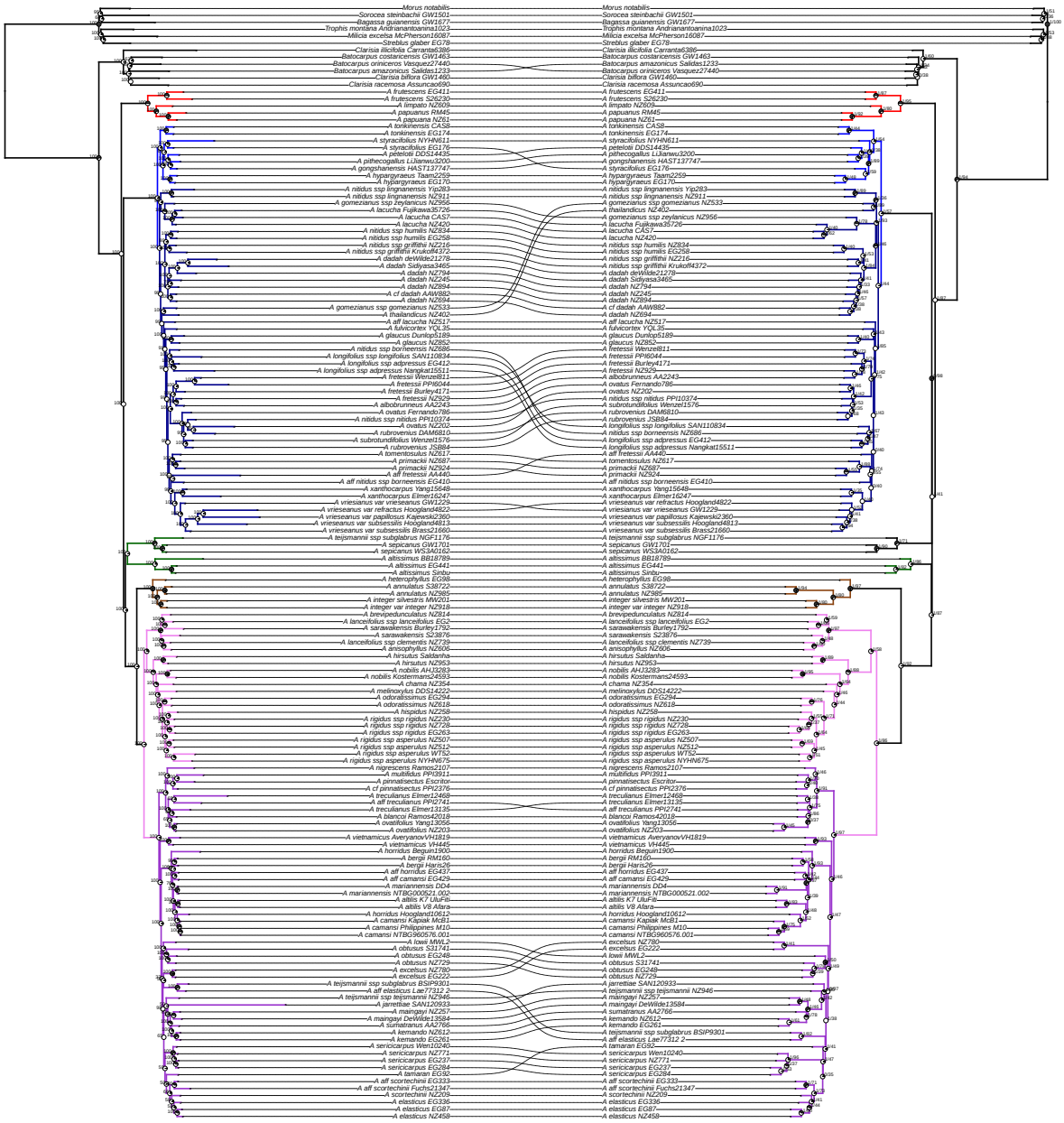


Figure 6. Comparison between the full-dataset (supercontigs for all genes) supermatrix and ASTRAL trees, showing moderate disagreement at shallow phylogenetic depths but complete agreement at deeper nodes. Left: maximum-likelihood tree based on all supercontigs, partitioned by gene, including all paralogs; all branch lengths are proportional to mean substitutions per site. Right: ASTRAL tree based on all supercontigs; internal

954 branch lengths are proportional to coalescent units; terminal branch lengths were arbitrarily
955 assigned to improve visualization. Pie charts at nodes represent the proportion of gene trees
956 supporting each split, and numbers represent bootstrap support
957