*Syst. Biol.* 0(0):1–18, 2020

DOI:10.1093/sysbio/syaa073

# Paralogs and Off-Target Sequences Improve Phylogenetic Resolution in a Densely Sampled Study of the Breadfruit Genus (*Artocarpus*, Moraceae)

ELLIOT M. GARDNER[1,2,3,4,5,*], MATTHEW G. JOHNSON[1,6], JOAN T. PEREIRA[7], AIDA SHAFREENA AHMAD PUAD[8], DEBY ARIFIANI[9], SAHROMI[10], NORMAN J. WICKETT[1,2] AND NYREE J.C. ZEREGA[1,2,*]

[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, IL 60532, USA;* [4]*Singapore Botanic Gardens, National Parks Board, 1 Cluny Road, 259569, Singapore;* [5]*Florida International University, Institute of Environment, 11200 SW 8th Street, OE 148 Miami, Florida 33199, USA;* [6]*Texas Tech University, Department of Biological Sciences, 2901 Main Street, Lubbock, TX 79409-3131, USA;* [7]*Forest Research Centre, Sabah Forestry Department, P.O. Box 1407, 90715 Sandakan, Sabah, Malaysia;* [8]*Faculty of Resource Science & Technology, Universiti Malaysia Sarawak, Kota Samarahan, Sarawak 94300, Malaysia;* [9]*Herbarium Bogoriense, Research Center for Biology, Indonesian Institute of Sciences, Cibinong, Jawa Barat, Indonesia; and* [10]*Center for Plant Conservation Botanic Gardens, Indonesian Institute Of Sciences, Bogor, Jawa Barat, Indonesia*

*Elliot M. Gardner and Matthew G. Johnson are co-first authors.*

**Correspondence to be sent to: Chicago Botanic Garden, Negaunee Institute for Plant Conservation Science and Action, 1000 Lake Cook Road, Glencoe, IL 60022, USA;*

*E-mail elliotgardner2012@u.northwestern.edu or n-zerega@northwestern.edu*

*Received 18 November 2019; reviews returned 31 August 2020; accepted 08 September 2020*
*Associate Editor: Michael Charleston*

*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. Although codon partitioning did not result in any substantial topological differences, the inclusion of flanking noncoding sequence in analyses significantly increased the resolution of gene trees. We also found that increasing the size of data sets 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. Although 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 data sets 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. [*Artocarpus*; Moraceae; noncoding sequences; phylogenomics; target enrichment.]

Reduced-representation methods such as target enrichment (HybSeq) have become important tools for phylogenetic studies, enabling high-throughput and cost-effective sequencing of hundreds of loci (Faircloth et al. 2012; Mandel et al. 2014; Weitemier et al. 2014). In this study, we employ HybSeq to investigate the breadfruit genus (*Artocarpus* J.R.Forst. & G.Forst., Moraceae), analyzing the utility of paralogs, partitioning, noncoding sequences, and herbarium specimens in reconstructing the most data-rich phylogeny of the genus to date.

HybSeq involves hybridizing a randomly sheared sequencing library to bait sequences, typically exons from one or more taxa within or near the target clade. Researchers have employed HybSeq in studies ranging from deep phylogenetics (Prum et al. 2015; Liu et al. 2019) to within-species phylogeography (Villaverde et al. 2018). It is particularly useful for recovering sequences from museum specimens, because target enrichment is suitable for very small DNA fragments and can help overcome the presence of contaminating nonendogenous DNA (Staats et al. 2013; Buerki and Baker 2016; Hart et al. 2016; Brewer et al. 2019). However, making the most of HybSeq data sets, which can comprise hundreds of thousands of characters, requires careful attention to assembly and analysis methods, particularly for degraded DNA from museum specimens. This particularly true because divergent analysis methods can sometimes lead to divergent topologies, all with apparently high statistical support.

The mechanics of HybSeq frequently result in the recovery of nontargeted sequences such as paralogs similar to the target sequences (Hart et al. 2016; Johnson et al. 2016, 2019; Liu et al. 2019) and noncoding sequences flanking the target sequences (e.g. Medina et al. 2019). Both were the case with HybSeq baits we previously developed for Moraceae phylogenetics (Gardner et al. 2016), many of which were represented as paralogous pairs in *Artocarpus* due to an ancient whole-genome duplication. In almost all cases, they were diverged enough to sort and analyze separately (Johnson et al. 2016). The same targets also typically recovered a several-hundred bp "splash zone" of flanking noncoding sequences (Johnson et al. 2016). The impact of off-target by-catch on phylogenetic reconstruction remains unclear but has the potential to greatly increase the number of phylogenetically informative genes. However, analysis of mixed coding and noncoding sequences can make it difficult to ensure that exons are aligned in frame, particularly when frameshifts are present (Ranwez et al. 2011), hampering partitioning of data sets