Low-Coverage, Whole-Genome Sequencing of Artocarpus camansi (Moraceae) for Phylogenetic Marker Development and Gene Discovery

Authors: Gardner, Elliot M., Johnson, Matthew G., Ragone, Diane, Wickett, Norman J., and Zerega, Nyree J. C.

Source: Applications in Plant Sciences, 4(7)

Published By: Botanical Society of America

URL: https://doi.org/10.3732/apps.1600017

Applications in Plant Sciences

Applications in Plant Sciences 2016 4(7): 1600017

Genomic Resources Note

Low-coverage, whole-genome sequencing of *Artocarpus camansi* (Moraceae) for phylogenetic marker development and gene discovery[1]

Elliot M. Gardner[2,3,5], Matthew G. Johnson[2], Diane Ragone[4], Norman J. Wickett[2,3], and Nyree J. C. Zerega[2,3]

[2]Department of Plant Science, Chicago Botanic Garden, 1000 Lake Cook Road, Glencoe, Illinois 60022 USA; [3]Plant Biology and Conservation, Northwestern University, 2205 Tech Drive, Hogan 2-144, Evanston, Illinois 60208 USA; and [4]Breadfruit Institute, National Tropical Botanical Garden, Kalaheo, Hawaii 96741 USA

- *Premise of the study:* We used moderately low-coverage (17×) whole-genome sequencing of *Artocarpus camansi* (Moraceae) to develop genomic resources for *Artocarpus* and Moraceae.
- *Methods and Results:* A de novo assembly of Illumina short reads (251,378,536 pairs, 2 × 100 bp) accounted for 93% of the predicted genome size. Predicted coding regions were used in a three-way orthology search with published genomes of *Morus notabilis* and *Cannabis sativa*. Phylogenetic markers for Moraceae were developed from 333 inferred single-copy exons. Ninety-eight putative MADS-box genes were identified. Analysis of all predicted coding regions resulted in preliminary annotation of 49,089 genes. An analysis of synonymous substitutions for pairs of orthologs (Ks analysis) in *M. notabilis* and *A. camansi* strongly suggested a lineage-specific whole-genome duplication in *Artocarpus*.
- *Conclusions:* This study substantially increases the genomic resources available for *Artocarpus* and Moraceae and demonstrates the value of low-coverage de novo assemblies for nonmodel organisms with moderately large genomes.

Key words: *Artocarpus camansi*; breadfruit; breadnut; genome; MADS box; Moraceae.

Artocarpus J. R. Forst. & G. Forst. (Moraceae) contains approximately 70 species of monoecious trees with a center of diversity in Malesia (Zerega et al., 2010). The genus includes several underutilized crops that can improve food security, most notably breadfruit (*A. altilis* (Parkinson) Fosberg), a long-lived perennial crop that is low input but high yielding (Jones et al., 2011). Other *Artocarpus* crops include the pantropically cultivated jackfruit (*A. heterophyllus* Lam.), crops of regional importance like cempedak (*A. integer* (Thunb.) Merr.) and terap (*A. odoratissimus* Blanco), and more than a dozen other species with edible fruits whose potential remains largely unexplored (Zerega et al., 2010). *Artocarpus camansi* Blanco (breadnut), native to New Guinea, is the diploid wild progenitor of breadfruit and is cultivated throughout the tropics for its large edible seeds (Zerega et al., 2005) (Fig. 1).

Existing genomic resources for breadfruit include nuclear (Witherup et al., 2013) and chloroplast (Gardner et al., 2015) microsatellites as well as transcriptomes of breadfruit and two wild relatives (Laricchia, 2014). In this study, we augment these resources with a low-coverage shotgun assembly of the *A. camansi* genome.

Recent studies focused on different taxonomic groups have used ultra-shallow sequencing (or "genome skimming")—with coverage of less than 1×—to assemble portions of genomes for annotation and marker discovery, particularly high-copy sequences such as organellar genomes and repetitive elements (Straub et al., 2011; Blischak et al., 2014). Here, we explore the utility of somewhat deeper but still shallow (17×) genome sequencing for de novo genome assembly and annotation with the goal of phylogenomic marker development, gene discovery, and the detection of whole-genome duplications.

METHODS AND RESULTS

The individual for sequencing was selected for the absence of heterozygosity at 19 nuclear microsatellite loci (Zerega et al., 2015), possibly due to centuries of inbreeding. The individual is the offspring of a tree planted in the Lancetilla Botanical Garden, Honduras. The Honduran tree likely descended from Caribbean material, which in turn was likely descended from seedlings in Mauritius that were collected in the Philippines by the French naturalist Pierre Sonnerat (1748–1814) (Ragone, 1997). Leaf tissue from *A. camansi* (living collection at McBryde Garden at the National Tropical Botanical Garden, Kalaheo, Hawaii: NTBG 960576.001; voucher: EG149 [CHIC]) was collected and dried on silica. Genomic DNA was extracted using the QIAGEN DNeasy Plant Mini Kit following the manufacturer's protocol (QIAGEN, Valencia, California, USA). Two Illumina TruSeq libraries were prepared: a paired-end library with a mean insert size of 180 bp and a mate-pair library with a mean insert size of 854 bp (Illumina, San Diego, California, USA). The paired-end library was sequenced in a single lane on an Illumina HiSeq 2000 (2 × 100 bp, paired–end), and the mate-pair library was sequenced in one-half lane. Library preparation and

[1]Manuscript received 12 February 2016; revision accepted 1 June 2016.

The authors thank S. Greenwald (Argonne National Laboratory) for library preparation and Illumina sequencing and A. Devault (MYcroarray) for assistance with bait design. This research was funded by National Science Foundation grants to N.J.C.Z. (DEB-0919119) and N.J.W. (DEB-1239992) and a grant to N.J.C.Z. from the Institute for Sustainability and Energy at Northwestern University.

[5]Author for correspondence: egardner@u.northwestern.edu

doi:10.3732/apps.1600017

Fig. 1. *Artocarpus camansi*, showing the individual used for sequencing. Photo by D. Ragone, NTBG accession 960576.001.

sequencing took place at the Next Generation Sequencing Core Facility, Argonne National Laboratory (Lemont, Illinois, USA).

***Assembly and bait sequence development*—**De-multiplexed sequences were quality trimmed (>Q20 in a 5-bp window) using Trimmomatic (Bolger et al., 2014) and assembled using Ray 2.3.1 with a k-mer size of 31 (Boisvert et al., 2010). We then conducted searches to identify sequences optimized for hybridization-based target enrichment (Hyb-Seq) baits. Hyb-Seq uses oligonucleotide fragments ("baits") to enrich DNA libraries for the desired portions of the genome ("targets") (Weitemier et al., 2014). Bait sequences for 333 phylogenetic markers were developed as follows: ESTScan 3.0.3 (Iseli et al., 1999;

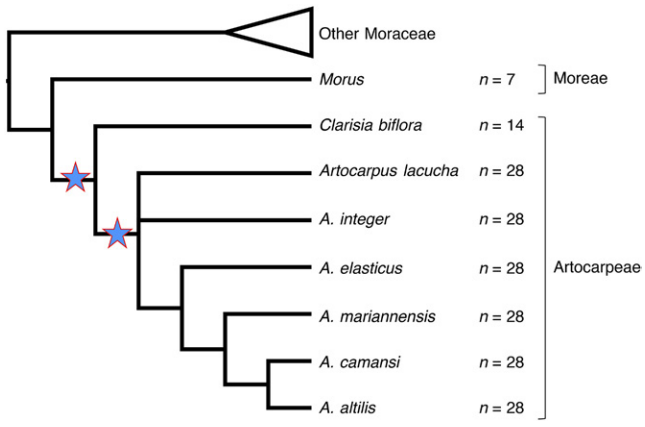

Fig. 2. Simplified Moraceae phylogeny based on Zerega et al. (2010) showing haploid chromosome numbers (Banerji and Hakim, 1954; Oginuma and Tobe, 1995; Ragone, 2001) and hypothesized whole-genome duplications (stars) in the tribe Artocarpeae.

Table 1. Sequencing statistics in *Artocarpus camansi*.

Sequencing statistics	Paired-end library	Mate-pair library
Mean insert size (bp)	180	854
Read pairs	182,485,953	68,892,583
Reads pairs surviving QC	161,937,040	53,824,945
GenBank SRA accession	SRR2910988	SRR2911031

http://sourceforge.net/projects/estscan/) was run on the *A. camansi* contigs and the *Cannabis sativa* L. transcriptome (van Bakel et al., 2011) to predict amino acid sequences. Homologous protein clusters (orthogroups) were circumscribed using reciprocal BLAST searches with these proteomes and the published proteome of *Morus notabilis* C. K. Schneid. (Moraceae) (He et al., 2013) using Proteinortho version 4 (Lechner et al., 2011). Orthogroups containing exactly one copy per species were subjected to further filtering as follows. Organellar sequences (determined using a BLASTN search against the nonredundant nucleotide [nr/nt] database) (Altschul et al., 1990) and predicted coding sequences residing on contigs smaller than 1000 bp were discarded. To ensure that baits would not hybridize to multiple regions of the genome, the remaining *Artocarpus* sequences were searched against the original *A. camansi* contigs using BLASTN, and any query sequence with more than one hit (>90% identity and >150 bp) was discarded. Finally, 333 predicted genes with exons between 500 and 2200 bp were selected. We retained only long exons to maximize the likelihood of hybridization success in divergent taxa, reasoning that longer exons would have a greater likelihood of containing at least a short span of conserved sequence that would hybridize across greater phylogenetic distances.

In addition to the markers targeted for their phylogenetic utility, target sequences for 125 additional genes were developed as follows. These sequences were selected based on predicted function only and were not screened for optimal hybridization characteristics such as long exons, high alignment specificity, or low copy number. Protein-coding regions of all scaffolds were predicted with AUGUSTUS (Stanke et al., 2008) using default parameters and allowing partial gene models. These predictions differed from the ESTScan predictions in their tendency to find even small, widely spaced, and partial exons, which is important for resequencing of functional genes, but not necessarily ideal for optimizing hybridization. Putative MADS-box genes were identified using a BLASTP search seeded with sequences from *Arabidopsis thaliana* (L.) Heynh. (Lamesch et al., 2012) and *M. notabilis* (He et al., 2013) containing "MADS" in their annotations.

An initial set of genes resulting from the BLASTP search (*E*-value cutoff 1e-10) was searched against all embryophyte genes in the National Center for Biotechnology Information nonredundant (NR) protein database, and those 98 sequences whose top BLASTP hit was a MADS-box gene were retained. To fill out the bait set, an additional 27 genes homologous to genes involved in biosynthesis of floral volatile compounds were extracted using the same search method, seeded with sequences from *Arabidopsis* and *Fragaria vesca* L. that are associated with KEGG pathway map00902 (monoterpenoid biosynthesis) (Kanehisa and Goto, 2000). These genes were of interest because of the role floral volatile compounds play in attracting pollinators in Moraceae (Yu et al., 2015). All 125 target sequences of functional interest were annotated with a BLASTP search against all plant proteins in the NR database, and ortholog hit ratios were calculated by dividing the alignment length by the target sequence length (O'Neil and Emrich, 2013). To tentatively classify the 98 MADS-box bait sequences, corresponding amino acid sequences were aligned using MAFFT (Katoh and Standley, 2013) to the 70 *Arabidopsis* MADS-box proteins chosen for the most recent comparative study using a member of Rosales (*Malus ×domestica* Borkh., domesticated apple) (Tian et al., 2015). We used amino acid rather than nucleotide sequences because of the great phylogenetic distances involved in aligning an entire ancient gene family. A maximum-likelihood tree was constructed with RAxML using the "PROTGAMMAAUTO" model setting (Stamatakis, 2006).

Table 2. Assembly statistics in *Artocarpus camansi.*

Assembly statistics	Ray assembly	Scaffolded with transcriptome
Scaffolds	401,889	388,956
Scaffold size (bp)	632,414,964	633,895,366
Scaffold N50	2426	2574
Contigs	415,187	414,958
Contig size (bp)	630,980,246	630,983,028
Contig N50	2315	2317

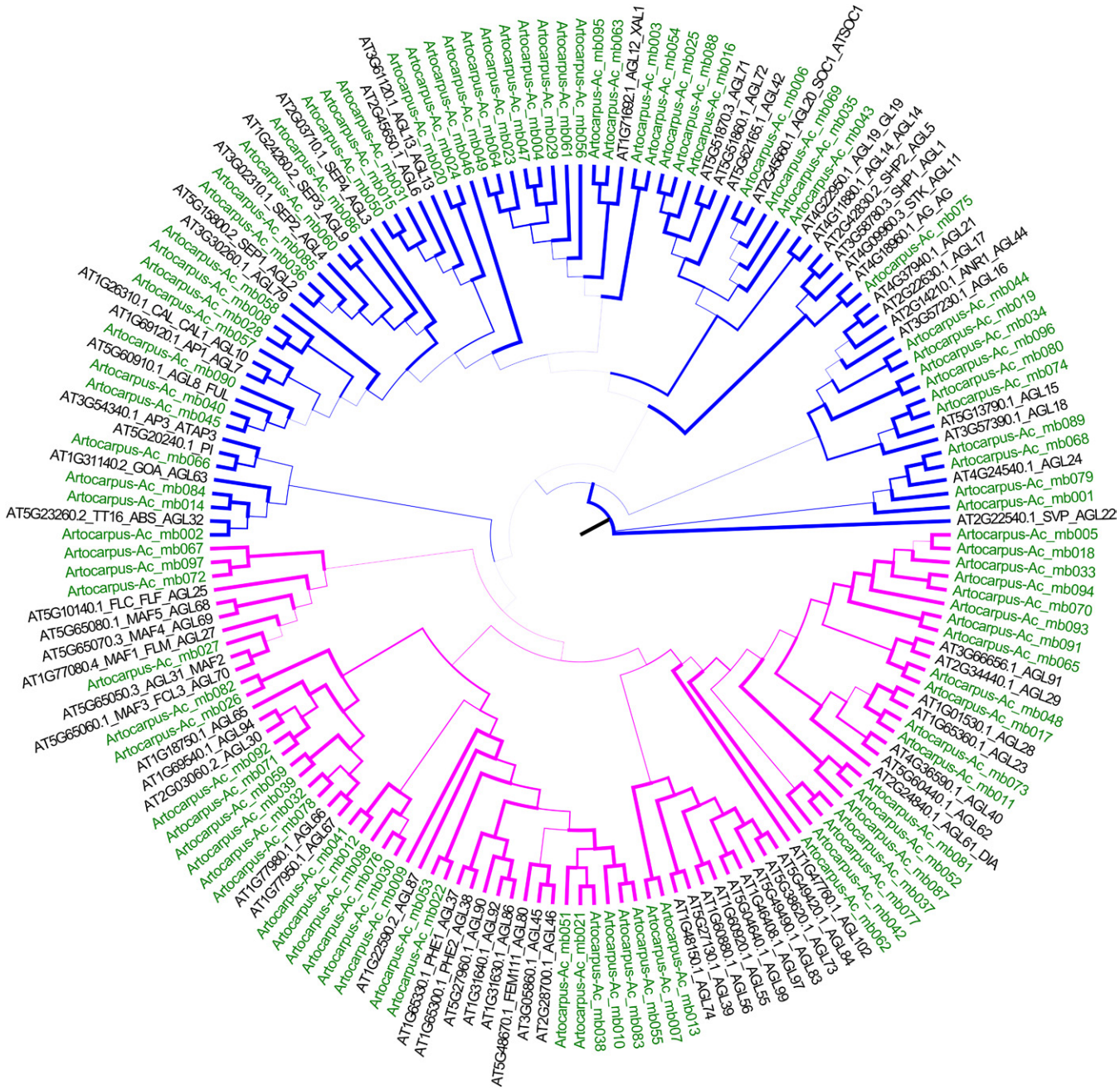

Fig. 3. Unrooted maximum-likelihood tree of *Arabidopsis* MADS-box genes (black) and putative *Artocarpus* MADS-box genes (green). Branches are color-coded based on *Arabidopsis* Type I (magenta) and Type II (blue) genes. Branch thickness is proportional to bootstrap support.

***RNA scaffolding, gene annotation, and Ks analysis*—**To scaffold the genome assembly further before a full genome annotation was performed, we used a transcriptome assembly of *A. camansi* (seed offspring of NTBG 000501.005, voucher: EG140 [CHIC], BioProject PRJNA311339) (Laricchia, 2014) in L_RNA_scaffolder (Xue et al., 2013). This final assembly was used for all further analyses. Gene prediction on the extended scaffolds was performed using AUGUSTUS (Stanke et al., 2008), guided by alignments from the transcriptome assembly, and limited to complete gene models only. Predicted amino acid sequences were annotated with Swiss-Prot hits and Gene Ontology (GO) terms using Trinotate (https://trinotate.github.io/). We determined orthology between the AUGUSTUS protein predictions and the published proteomes of *M. notabilis* and *C. sativa* with Proteinortho (Lechner et al., 2011), using default settings. *Artocarpus* proteins were reduced to a nonredundant set using CD-HIT-EST (Fu et al., 2012) to remove sequences that were 98% similar at the nucleotide level along 90% of their length. Chromosome counts support a hypothesis of two whole-genome duplications within Artocarpeae: the one at the root of the tribe, evidenced by a haploid chromosome number of 14 in *Clarisia biflora* Ruiz & Pav. (Oginuma and Tobe, 1995), and another in the genus *Artocarpus*, which has a haploid number of 28 (Banerji and Hakim, 1954; Oginuma and Tobe, 1995; Ragone, 2001). By contrast, *Morus* has a base haploid number of 7, although a polyploid series exists within that genus as well (He et al., 2013) (Fig. 2). To determine the age of possible whole-genome duplication (WGD) in *Artocarpus*, we selected Proteinortho orthogroups that were single copy in *Morus* and *Cannabis*, but multicopy in *Artocarpus*. We generated pairwise alignments of *Artocarpus* protein sequences (paralogs) in each of these orthogroups using MAFFT, and back-translated the sequences to nucleotides using PAL2NAL (Suyama et al., 2006). We calculated the synonymous substitution rate (Ks) for each paralog pair with KaKs_Calculator (Zhang et al., 2006) using the GY

method (also known as F3 × 4) (Goldman and Yang, 1994). We also used this procedure to calculate Ks for orthogroups that were single copy in all three proteomes to generate three ortholog Ks distributions: *Morus-Artocarpus*, *Morus-Cannabis*, and *Cannabis-Artocarpus*.

***Sequencing and assembly results*—**Sequencing and assembly statistics appear in Tables 1 and 2. Based on the k-mer counting method, the Ray assembler predicted a genome size of 669 Mb, with a peak coverage of 17×. The Ray assembly was 631 Mb (93% of the predicted size by the k-mer method implemented in Ray), and the scaffold N50 was 2426 bp. After additional scaffolding with L_RNA_scaffolder, the scaffold N50 increased to 2574, and the number of scaffolds dropped from 401,899 to 388,956 (Table 2).

***Identification of target sequences*—**Proteinortho found 2041 putative single-copy orthogroups with exactly one copy in each of the three species included that were nonorganellar and were present on contigs longer than 1000 bp. Of the 617 orthogroups with only a single BLASTN hit in the original *A. camansi* assembly hit (>90% identity and >150 bp), the 333 selected for bait design were between 504 and 2166 bp, and these totaled 354,171 bp. Comparison of these exons with full gene predictions by AUGUSTUS in IGV revealed that ESTScan tended to find single-exon genes or the longest exon in multiexon genes, which was expected to increase the efficiency of hybridization. GenBank accession numbers, *Artocarpus* scaffold coordinates, and orthology assignments to *M. notabilis* and *C. sativa* are presented in Appendix S1. The maximum-likelihood tree (Fig. 3) indicates that of the 98 putative MADS-box sequences, 51 are tentatively assignable to domain MIKC (Type 2) and 47 are tentatively assignable to domain M (Type 1). However, these assignments should be treated with caution due to the fragmentary nature of the assembly. Exons for the putative MADS-box genes and the additional 27 terpenoid synthesis genes totaled 101,871 bp. The median ortholog hit ratio, calculated against the top BLASTP hit, was 0.60, indicating that most sequences likely contained the majority of a gene; 16 had an ortholog hit ratio of 1, indicating complete coding sequences. GenBank accession numbers, *Artocarpus* scaffold coordinates, and the best BLASTP hit are presented in Appendix S2. Fine-scale name assignments based on BLASTP hits should be treated with caution, particularly for the MADS-box genes, as they do not always agree with the phylogenetic placement in the maximum-likelihood tree. The 458 *Artocarpus* sequences identified here, combined with *Morus* orthologs for the 333 phylogenetic markers, maximized the available target space for a bait set made up of 20,000 120-mer baits with 3× tiling (MYcroarray, Ann Arbor, Michigan, USA). The use of these bait sequences to capture sequences from taxa representing all Moraceae tribes, including differences in hybridization efficiency at various phylogenetic distances, and between the 333 phylogenetic markers and the other genes, is reported in Johnson et al. (2016).

***Gene annotation and Ks analysis*—**AUGUSTUS predicted 83,061 gene sequences on the extended scaffolds; of these, Trinotate reported annotations for 49,089 sequences, including 22,865 sequences annotated with 1226 unique GO terms. A search for reciprocal best BLASTP hits between the predicted proteins for *A. camansi* and *M. notabilis* found 14,089 orthogroups collectively containing 15,634 *Artocarpus* genes and 14,506 *Morus* genes, which is 54% of the *Morus* total. Using the AUGUSTUS protein predictions, the *M. notabilis* proteome, and the *C. sativa* proteome, we generated 18,657 orthogroups using Proteinortho, 10,925 of which contained proteins from all three proteomes, and 8,539 were single copy in all three genomes. There were 1391 orthogroups containing multiple sequences from *Artocarpus* but only one sequence from each of the published genomes (many-to-one orthogroups). The distribution of synonymous substitutions (Ks) between pairs of *Artocarpus* paralogs in the many-to-one orthogroups reflects evidence of a possible lineage-specific WGD in *A. camansi* (Fig. 4). We then compared the Ks distribution of *Artocarpus* paralogs in many-to-one orthogroups to the distribution of Ks between pairs of orthologs in the single-copy orthogroups generated by Proteinortho. The *Artocarpus* paralog distribution indicates WGD in *Artocarpus* is more recent than the divergence of *Artocarpus* from *M. notabilis*, a member of the sister tribe Moreae (Fig. 4). This is consistent with a hypothesis of WGD events in the tribe Artocarpeae based on published chromosome counts (Fig. 2, discussed above). However, as confirmed chromosome counts exist for only a few of the ca. 70 *Artocarpus* species, more work is required to arrive at a firm conclusion regarding the precise phylogenetic position of the duplication within the genus.

The raw reads and the Ray assembly were deposited in GenBank under BioProject PRJNA301299. Scaffolds less than 200 bp (n = 5813) were omitted from the GenBank accession, because NCBI WGS does not accept scaffolds shorter than 200 bp. Both draft assemblies, bait sequences, and full-length targets described here have been archived in the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.3293r; Gardner et al., 2016). They are also available on the *Artocarpus* Genomics website, which will host any future updates (http://sites.northwestern.edu/zerega-lab/research/artocarpus-genomics/).

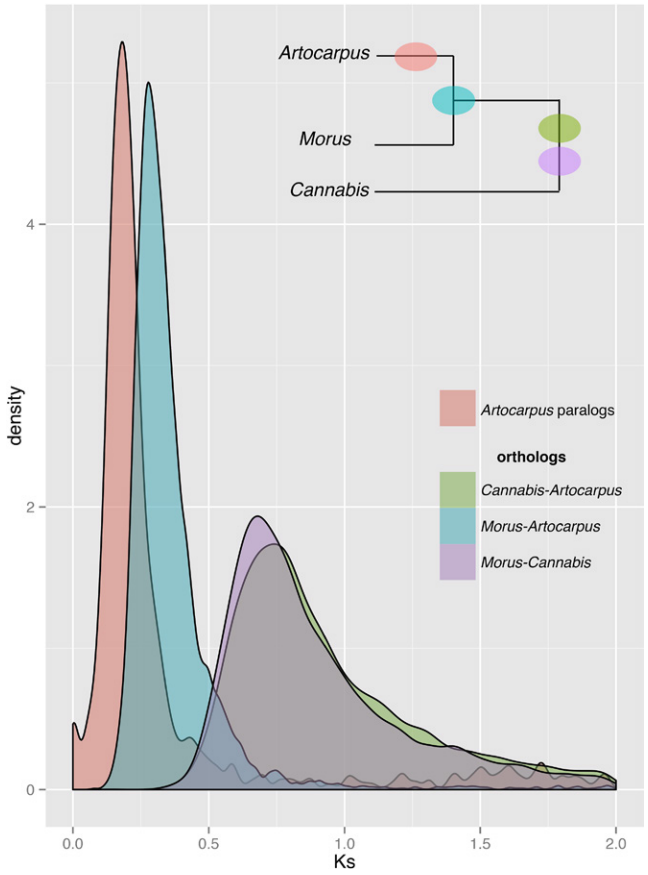

Fig. 4. Distribution of synonymous substitution rates (Ks) for paralog pairs in *Artocarpus*, and orthologs between *Artocarpus*, *Morus*, and *Cannabis* proteomes, demonstrating a possible lineage-specific whole-genome duplication in *Artocarpus*. Paralogs are from orthogroups reconstructed to be multicopy in *Artocarpus* but single copy in both *Morus* and *Cannabis*. Orthologs between pairs of species are from orthogroups reconstructed to be single copy in all three proteomes.

CONCLUSIONS

This study constitutes a substantial increase in the genomic resources available for *Artocarpus* in particular and Moraceae in general. The 333 phylogenetic markers will enable large-scale phylogenomic studies within the family using a Hyb-Seq (target enrichment) approach. Because sequence capture is anchored by conserved exons but can extend into flanking noncoding regions if read lengths are sufficient (Johnson et al., 2016), these markers can be employed at both deep and shallow phylogenetic scales. The gene predictions provide a resource that can be mined to explore gene families; for example, the MADS-box sequences identified provide a starting point for characterizing this important gene family in *Artocarpus*. Finally, the shotgun assembly presented here, while fragmentary, covers an estimated 93% of the genome, and represents a first step toward a reference genome assembly for *Artocarpus*. Thus the study demonstrates how existing reference genomes (*Morus*, *Cannabis*, and *Arabidopsis*)

can be leveraged to increase the value of de novo shotgun assemblies from shallow genome sequencing, enabling the inexpensive development of genome-level resources for nonmodel organisms with moderately large genomes.

LITERATURE CITED

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215: 403–410.

Banerji, I., and A. Hakim. 1954. A contribution to the life-history of *Artocarpus lakoocha* Roxb. *Proceedings of the Indian Academy of Sciences, Section B* 39: 128–132.

Blischak, P. D., A. J. Wenzel, and A. D. Wolfe. 2014. Gene prediction and annotation in *Penstemon* (Plantaginaceae): A workflow for marker development from extremely low-coverage genome sequencing. *Applications in Plant Sciences* 2: 1400044.

Boisvert, S., F. Laviolette, and J. Corbeil. 2010. Ray: Simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *Journal of Computational Biology* 17: 1519–1533.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)* 30: 2114–2120.

Fu, L., B. Niu, Z. Zhu, S. Wu, and W. Li. 2012. CD-HIT: Accelerated for clustering the next generation sequencing data. *Bioinformatics (Oxford, England)* 28: 3150–3152.

Gardner, E. M., K. M. Laricchia, M. Murphy, D. Ragone, B. E. Scheffler, S. Simpson, E. W. Williams, and N. J. C. Zerega. 2015. Chloroplast microsatellite markers for *Artocarpus* (Moraceae) developed from transcriptome sequences. *Applications in Plant Sciences* 3(9): 1500049.

Gardner, E. M., M. G. Johnson, D. Ragone, N. J. Wickett, and N. J. C. Zerega. 2016. Data from: Low-coverage, whole-genome sequencing of *Artocarpus camansi* (Moraceae) for phylogenetic marker development and gene discovery. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.3293r

Goldman, N., and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* 11: 725–736.

He, N., C. Zhang, X. Qi, S. Zhao, and Y. Tao. 2013. Draft genome sequence of the mulberry tree *Morus notabilis*. *Nature Communications* 4: 2445.

Iseli, C., C. V. Jongeneel, and P. Bucher. 1999. ESTScan: A program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *In* T. Lengauer, R. Schneider, P. Bork, D. Brutlag, J. Glasgow, H.-W. Mewes, and R. Zimmer [eds.], Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology, 138–148. AAAI Press, Palo Alto, California, USA.

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(7): 1600016.

Jones, A. M. P., D. Ragone, N. G. Tavana, D. W. Bernotas, and S. J. Murch. 2011. Beyond the bounty: Breadfruit (*Artocarpus altilis*) for food security and novel foods in the 21st century. *Ethnobotany Research and Applications* 9: 129–149.

Kanehisa, M., and S. Goto. 2000. KEGG: Kyoto Encyclopaedia of Genes and Genomes. *Nucleic Acids Research* 28: 27–30.

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.

Lamesch, P., T. Z. Berardini, D. Li, D. Swarbreck, C. Wilks, R. Sasidharan, R. Muller, et al. 2012. The *Arabidopsis* Information Resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Research* 40: D1202–D1210.

Laricchia, K. 2014. Transcriptome analysis of domesticated breadfruit and its wild relatives. Master's thesis, Northwestern University, Evanston, Illinois, USA.

Lechner, M., S. Findeiss, L. Steiner, M. Marz, P. F. Stadler, and S. J. Prohaska. 2011. Proteinortho: Detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics* 12: 124.

O'Neil, S. T., and S. J. Emrich. 2013. Assessing de novo transcriptome assembly metrics for consistency and utility. *BMC Genomics* 14: 465.

Oginuma, K., and H. Tobe. 1995. Karyomorphology of some Moraceae and Cecropiaceae (Urticales). *Journal of Plant Research* 108: 313–326.

Ragone, D. 1997. Breadfruit, *Artocarpus altilis* (Parkinson) Fosberg. Promoting the conservation and use of underutilized and neglected crops. IPIGRI, Rome, Italy.

Ragone, D. 2001. Chromosome numbers and pollen stainability of three species of Pacific Island breadfruit (*Artocarpus*, Moraceae). *American Journal of Botany* 88: 693–696.

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

Stanke, M., M. Diekhans, R. Baertsch, and D. Haussler. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics (Oxford, England)* 24: 637–644.

Straub, S. C. K., M. Fishbein, T. Livshultz, Z. Foster, M. Parks, K. Weitemier, R. C. Cronn, and A. Liston. 2011. Building a model: Developing genomic resources for common milkweed (*Asclepias syriaca*) with low coverage genome sequencing. *BMC Genomics* 12: 211.

Suyama, M., D. Torrents, and P. Bork. 2006. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research* 34: W609–W612.

Tian, Y., Q. Dong, Z. Ji, F. Chi, P. Cong, and Z. Zhou. 2015. Genome-wide identification and analysis of the MADS-box gene family in apple. *Gene* 555: 277–290.

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

Weitemier, K., S. C. K. Straub, R. C. Cronn, M. Fishbein, R. Schmickl, A. McDonnell, and A. Liston. 2014. Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences* 2(9): 1400042.

Witherup, C., D. Ragone, T. Wiesner-Hanks, B. Irish, B. Scheffler, S. Simpson, F. Zee, et al. 2013. Development of microsatellite loci in *Artocarpus altilis* (Moraceae) and cross-amplification in congeneric species. *Applications in Plant Sciences* 1(7): 1200423.

Xue, W., J.-T. Li, Y.-P. Zhu, G.-Y. Hou, X.-F. Kong, Y.-Y. Kuang, and X.-W. Sun. 2013. L_RNA_scaffolder: Scaffolding genomes with transcripts. *BMC Genomics* 14: 604.

Yu, H., J. D. Nason, L. Zhang, L. Zheng, W. Wu, and X. Ge. 2015. De novo transcriptome sequencing in *Ficus hirta* Vahl. (Moraceae) to investigate gene regulation involved in the biosynthesis of pollinator attracting volatiles. *Tree Genetics & Genomes* 11: 91.

Zerega, N. J. C., D. Ragone, T. Motley, and W. Zomlefer. 2005. Systematics and species limits of breadfruit (*Artocarpus*, Moraceae). *Systematic Botany* 30: 603–615.

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

Zerega, N., T. Wiesner-Hanks, D. Ragone, B. Irish, B. Scheffler, S. Simpson, and F. Zee. 2015. Diversity in the breadfruit complex (*Artocarpus*, Moraceae): Genetic characterization of critical germplasm. *Tree Genetics & Genomes* 11: 4.

Zhang, Z., J. Li, X. Q. Zhao, J. Wang, G. K. S. Wong, and J. Yu. 2006. KaKs_Calculator: Calculating Ka and Ks through model selection and model averaging. *Genomics, Proteomics & Bioinformatics* 4: 259–263.