Data in Brief 7 (2016) 532–536

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Data supporting the nuclear phylogenomics of the palm subfamily Arecoideae (Arecaceae)

Jason R. Comer [a,*], Wendy B. Zomlefer [a], Craig F. Barrett [b,1], Dennis Wm. Stevenson [c], Karolina Heyduk [a], James H. Leebens-Mack [a]

[a] University of Georgia, Department of Plant Biology, Athens, GA 30602–7271, USA
[b] California State University, Los Angeles, Department of Biological Sciences, Los Angeles, CA 90032–8201, USA
[c] New York Botanical Garden, Bronx, NY 10458–5126, USA

ARTICLE INFO

Article history:
Received 4 January 2016
Received in revised form
14 February 2016
Accepted 22 February 2016
Available online 2 March 2016

Keywords:
Ancestral area
Arecaceae
Arecoideae
Coalescent
Nuclear phylogeny
Targeted sequencing

ABSTRACT

This data article provides data and supplemental materials referenced in "Nuclear phylogenomics of the palm subfamily Arecoideae (Arecaceae)" (Comer et al., 2016) [1]. Raw sequence reads generated for this study are available through the Sequence Read Archive (SRA Study Accession: SRP061467). An aligned supermatrix of 168 nuclear genes for 35 taxa (34 palms and one outgroup taxon) is provided. Also provided are individual maximum likelihood gene trees used for the coalescent based analyses, output from the maximum parsimony analyses, and two figures.

Specifications Table

Subject area	Biology, Genetics and Genomics
More specific subject area	Phylogenetics and Phylogenomics

DOI of original article: http://dx.doi.org/10.1016/j.ympev.2015.12.015
* Corresponding author.
E-mail addresses: jcomer@uga.edu (J.R. Comer), wendyz@uga.edu (W.B. Zomlefer), craigbarrett.barrett@mail.wvu.edu (C.F. Barrett), dws@nybg.org (D.Wm. Stevenson), heyduk@uga.edu (K. Heyduk), jleebensmack@uga.edu (J.H. Leebens-Mack).
[1] Present address: Division of Plant and Soil Sciences, West Virginia University, Morgantown, WV 26506, USA

http://dx.doi.org/10.1016/j.dib.2016.02.063
2352-3409/

Type of data	*Sequence alignment, analysis output file, and figures*
How data was acquired	*Hybrid gene capture and Illumina MiSeq sequencing.*
Data format	*Raw and analyzed.*
Experimental factors	*Hybrid gene capture on total genomic DNA, following the protocol of Heyduk et al. [2] and Comer et al. [3].*
Experimental features	*Following hybridization and sequencing, 168 nuclear genes (for 35 taxa) were used for phylogenetic analyses.*
Data source location	*Newly sampled taxa for this dataset were collected from Cameroon, Florida, Ghana, and Thailand. See also Appendix A in Comer et al. [1].*
Data accessibility	*Data is within this article. For raw sequence reads see SRA Study Accession: SRP061467.*

Value of the data

- Provides a dataset of 168 nuclear genes for 34 palm taxa and one outgroup taxon.
- Provides a nuclear phylogeny for the palm family from the largest dataset to date.
- Provides a foundational dataset for future phylogenomic studies of palms.

1. Data

The dataset shared here consists of the 168 aligned nuclear gene supermatrix (Supplementary material 1) used in Comer et al. [1]. Also shared within this article are supporting material referenced in Comer et al. [1] (Supplementary material 2–4 and Figs. 1 and 2).

2. Experimental design, materials and methods

2.1. Taxon sampling and hybrid gene capture

Thirty-four species were sampled, representing the five palm subfamilies and the 14 tribes of subfamily Arecoideae (see Comer et al. [1] Appendix A). Total genomic DNA was sheared with a Covaris sonicator (Woburn, MA, USA) to an appropriate size then used for Illumina library construction (see also Comer et al. [1,3] and Heyduk et al. [2]). Resulting genomic libraries were enriched for target nuclear exons through hybridization to RNA baits (MYcoarray, Ann Arbor, Michigan, USA) [2–4]. Hybridization reactions were pooled for paired-end sequencing on the Illumina MiSeq platform [3].

2.2. Assembly

Sequence reads were demultiplexed, quality trimmed from the 3′ ends, and filtered [1–3]. The *de novo* assembler Trinity v. 2.06 [5] was used to assemble the cleaned reads, and CAP3 v. 102011 [6] was used to collapse assembled contigs [1]. Assembled contigs with segments matching the target exons were identified using BLAST (Basic Local Alignment Search Tool; Expect value 1×10^{-20}; [7]). Following Heyduk et al. [2], duplicate contigs were removed to reduce the potential for paralogy (see Fig. 2b in Comer et al. [1]). Exons from the same gene were concatenated into super scaffolds. For summary statistics see Table 2 in Comer et al. [1].

Assembled genes were aligned using PRANK v. 100802 [8], and Gblocks v. 0.91b [9] was used to filter poorly aligned and non-conserved regions [1]. Genes were excluded if a significant amount of data was missing or if the aligned gene exhibited an average pair-wise genetic distance of more than 0.15 [1]. Scripts used for this study's assembly pipeline can be found at: https://github.com/kheyduk/reads2trees.

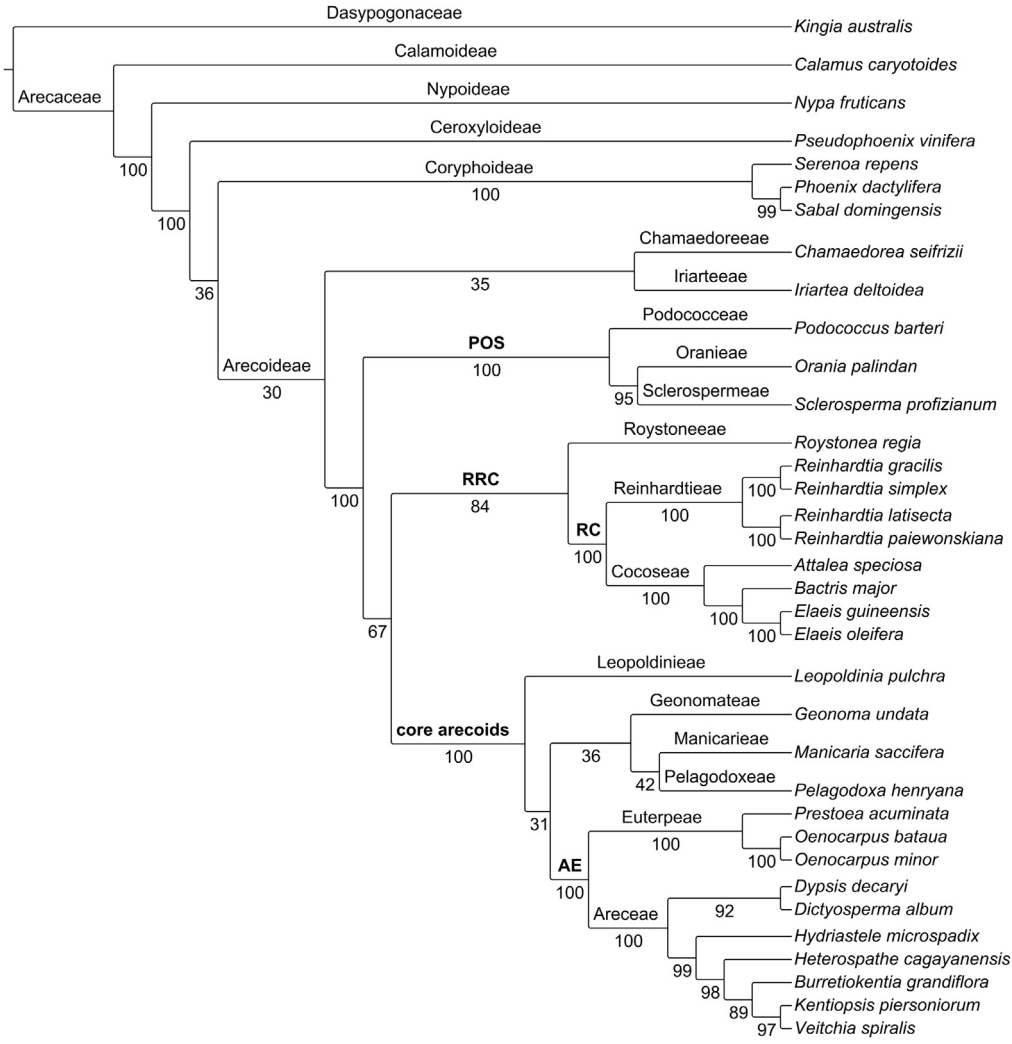

Fig. 1. Species tree (most parsimonious) from the MP concatenated analysis of the 168 nuclear genes. Labels above the branches=family, subfamily, tribe, and major clade (boldface font); labels below branches=bootstrap support. Major clades: AE (Areceae+Euterpeae), core arecoids (Areceae, Euterpeae, Geonomateae, Leopoldinieae, Manicarieae, and Pelagodoxeae), POS (Podococceae, Oranieae, and Sclerospermeae), RC (Reinhardtieae+Cocoseae), and RRC (Roystoneeae, Reinhardtieae, and Cocoseae).

2.3. Phylogenetic reconstruction

Phylogenetic analyses were performed using supermatrix and coalescence-based species tree estimation approaches utilizing the 168 nuclear gene dataset presented here (Supplementary material 1). For the maximum parsimony, aligned genes were concatenated into a single supermatrix alignment (Supplementary material 1) and the TNT v. 1.1 (Tree Analysis Using New Technology, Willi Hennig Society edition; [10,11]) "one-shot" analysis script (consecutively ran random addition sequences, TBR, sectorial searches, and tree fusing each iteration for 20 iterations, 100 random addition replications and 1000 standard bootstrap replicates) was used for phylogenetic reconstruction (Fig. 1 and Supplementary material 2). ASTRAL v. 4.7.8, a coalescent based species tree estimation method, was used to estimate the species tree [12] from individual gene trees and bootstrap replicates estimated with

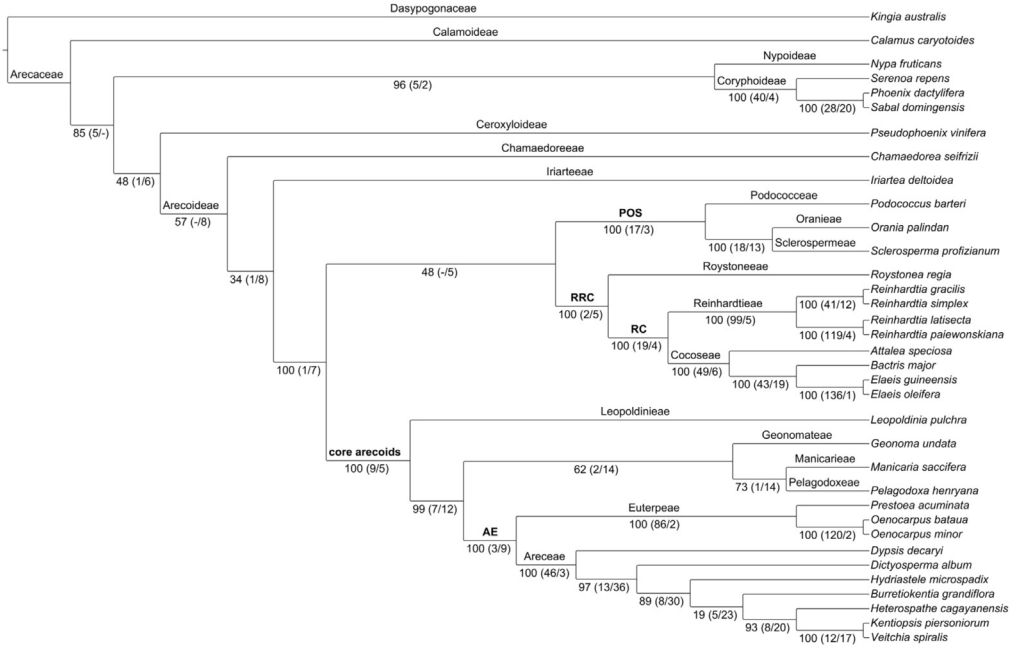

Fig. 2. Species tree from the ASTRAL analysis of the best gene trees of the 168 nuclear genes. Labels above the branches=family, subfamily, tribe, and major clade (boldface font); labels below branches=bootstrap support; numbers in parentheses=gene trees supporting (monophyletic) or rejecting (polyphyletic) the clade with a bootstrap value ≥ 75; a dash (–) indicates no genes trees with a bootstrap value of ≥ 75. Major clades: AE (Areceae+Euterpeae), core arecoids (Areceae, Euterpeae, Geonomateae, Leopoldinieae, Manicarieae, and Pelagodoxeae), POS (Podococceae, Oranieae, and Sclerospermeae), RC (Reinhardtieae+Cocoseae), and RRC (Roystoneeae, Reinhardtieae, and Cocoseae).

RAxML (GTRGAMMA, '-f a', and 500 bootstrap replicates; Supplementary material 3 and 4) [13–15]. We used the ASTRAL's heuristic version to implement a multi-locus bootstrapping analysis for both the ML best scoring gene trees (Fig. 2) and the ML bootstrap replicates (Fig. 3 in Comer et al. [1]).

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2016.02.063.

References

[1] J.R. Comer, W.B. Zomlefer, C.F. Barrett, D.W. Stevenson, K. Heyduk, J. Leebens-Mack, Nuclear phylogenomics of the palm subfamily Arecoideae (Arecaceae), Mol. Phylogenet. Evol. 97 (2016) 42. http://dx.doi.org/10.1016/j.ympev.2015.12.015.
[2] K. Heyduk, D.W. Trapnell, C.F. Barrett, J. Leebens-Mack, Phylogenomic analyses of *Sabal* (Arecaceae) species relationships using targeted sequence capture, Bot. J. Linn. Soc. (2015), http://dx.doi.org/10.1111/bij.12551.
[3] J.R. Comer, W.B. Zomlefer, C.F. Barrett, J.I. Davis, D.W. Stevenson, K. Heyduk, J. Leebens-Mack, Resolving relationships within the palm subfamily Arecoideae (Arecaceae) using next-gen derived plastid sequences, Am. J. Bot. 102 (2015) 888–899. http://dx.doi.org/10.3732/ajb.1500057.
[4] S. Fisher, A. Barry, J. Abreu, B. Minie, J. Nolan, T. Delorey, G. Young, T. Fennell, A. Allen, L. Ambrogio, et al., A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries, Genome Biol. 12 (2011) R1.
[5] M.G. Grabherr, B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Full-length transcriptome assembly from RNA-Seq data without a reference genome, Nat. Biotechnol. 29 (2011) 644–652.
[6] X. Huang, A. Madan, CAP3: a DNA sequence assembly program, Genome Res. 9 (1999) 868–877.
[7] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.

[8] A. Löytynoja, N. Goldman, An algorithm for progressive multiple alignment of sequences with insertions, Proc. Natl. Acad. Sci. USA 102 (2005) 10557–10562. http://dx.doi.org/10.1073/pnas.0409137102.
[9] J. Castresana, Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis, Mol. Biol. Evol. 17 (2000) 540–552.
[10] P.A. Goloboff, Analyzing large data sets in reasonable times: solutions for composite optima, Cladistics 15 (1999) 415–428.
[11] P.A. Goloboff, J.S. Farris, K.C. Nixon, TNT, a free program for phylogenetic analysis, Cladistics 24 (2008) 774–786.
[12] S. Mirarab, R. Reaz, M.S. Bayzid, T. Zimmermann, M.S. Swenson, T. Warnow, ASTRAL: genome-scale coalescent-based species tree estimation, Bioinformatics 30 (2014) i541–i548. http://dx.doi.org/10.1093/bioinformatics/btu462.
[13] A. Stamatakis, RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models, Bioinformatics 22 (2006) 2688–2690. http://dx.doi.org/10.1093/bioinformatics/btl446.
[14] A. Stamatakis, RAxML Version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies, Bioinformatics 30 (2014) 1312–1313. http://dx.doi.org/10.1093/bioinformatics/btu033.
[15] A. Stamatakis, F. Blagojevic, D.S. Nikolopoulos, C.D. Antonopoulos, Exploring new search algorithms and hardware for phylogenetics: RAxML meets the IBM cell, J. VLSI Sign. Process. Syst. Sign. Im. 48 (2007) 271–286. http://dx.doi.org/10.1007/s11265-007-0067-4.