The Linnean Society of London BIOLOGICAL Journal of the Linnean Society

Biological Journal of the Linnean Society, 2015, ••, ••–••. With 4 figures.

Phylogenomic analyses of species relationships in the genus *Sabal* (Arecaceae) using targeted sequence capture

KAROLINA HEYDUK[1]*, DORSET W. TRAPNELL[1], CRAIG F. BARRETT[2] and JIM LEEBENS-MACK[1]

[1]*Department of Plant Biology, University of Georgia, Athens, GA, 30602, USA*
[2]*Department of Biological Sciences, California State University, Los Angeles, CA, 90032, USA*

Received 30 December 2014; revised 2 March 2015; accepted for publication 6 March 2015

With the increasing availability of high-throughput sequencing, phylogenetic analyses are no longer constrained by the limited availability of a few loci. Here, we describe a sequence capture methodology, which we used to collect data for analyses of diversification within *Sabal* (Arecaceae), a palm genus native to the south-eastern USA, Caribbean, Bermuda and Central America. RNA probes were developed and used to enrich DNA samples for putatively low copy nuclear genes and the plastomes for all *Sabal* species and two outgroup species. Sequence data were generated on an Illumina MiSeq sequencer and target sequences were assembled using custom workflows. Both coalescence and supermatrix analyses of 133 nuclear genes were used to estimate species trees relationships. Plastid genomes were also analysed, yielding generally poor resolution with regard to species relationships. Species relationships described in both nuclear gene and plastome sequences largely reflect the biogeography of the group and, to a lesser extent, previous morphology-based hypotheses. Beyond the biological implications, this research validates a high-throughput methodology for generating a large number of genes for coalescence-based phylogenetic analyses in plant lineages. © 2015 The Linnean Society of London, *Biological Journal of the Linnean Society*, 2015, **00**: 000–000.

ADDITIONAL KEYWORDS: coalescence – gene trees – Palmae – palms – plastome.

INTRODUCTION

The palm family (Arecaceae) represents one of the most diverse monocot lineages, with approximately 2600 species in over 188 genera. With growth habits ranging from shrubby to arborescent, palms have colonized much of the tropics and subtropics and display many instances of endemism and island dispersal (Bacon, Baker & Simmons, 2012). Phylogenetic studies of the palms have often been hindered by low rates of molecular evolution and a relatively high incidence of homoplasious morphological characters (Baker, Dransfield & Hedderson, 2000; Hahn, 2002; Smith & Donoghue, 2008). Phylogenetic investigations of speciation processes typically involve analysis of intrageneric relationships, but resolving species-level relationships may be particularly difficult with the low rates of molecular evolution found in the Arecaceae.

Sabal Adans. (subfamily Coryphoideae, tribe Palmae, Arecaceae) is found throughout the south-eastern United States, Caribbean and Bermuda, as well as central and western Mexico and into northern South America (Dransfield *et al.*, 2008; Baker *et al.*, 2009). Discontinuity in the geographical distribution of *Sabal* species suggests that both vicariance and dispersal may have contributed to diversification within the genus. It is described as somewhat weedy and is used horticulturally and for basket fibres in the south-eastern USA (Zona, 1990). The genus is well defined by morphological characters such as unarmed petioles with a triangular cleft at the base, strongly costapalmate fronds, small solitary hermaphroditic flowers and small black fruit typically with a single seed (Zona, 1990). Within the genus, however, resolution of species relationships has been difficult due to a paucity of informative characters. Adding to the challenge, many morphological traits can vary widely within a species (Zona, 1990). To date, there has been very little assessment of molecular variation within and among *Sabal* species (but see Goldman *et al.*,

*Corresponding author. E-mail: kheyduk@plantbio.uga.edu

2011). Zona's cladogram, based on 22 morphological characters, generally groups species by geographical proximity; the species within the USA form a clade and a large group of Mexican/Yucatan/South American species are placed sister to the Caribbean and Bermudan species.

As many as three species are endemic to the Caribbean islands (Zona, 1990). The most geographically isolated species is *Sabal bermudana*, which is restricted to the island of Bermuda, 1030 km from the USA coast. Zona suggested that bird-mediated seed dispersal was responsible for the occurrence of *Sabal* on Bermuda and the Caribbean islands; while the seeds of some *Sabal* species were found to be viable after prolonged floatation experiments, there is no experimental evidence of hydrochory (Zona, 1990). Of all the island species, *S. bermudana*'s extreme isolation requires the furthest long-distance dispersal event, probably mediated by birds drawn to the sweet pericarp (Zona, 1990).

Here we take a phylogenomic approach utilizing targeted sequence capture (Faircloth *et al.*, 2012; Grover, Salmon & Wendel, 2012; Lemmon, Emme & Lemmon, 2012; McCormack *et al.*, 2013; Mandel *et al.*, 2014; Weitemier *et al.*, 2014) to test the phylogenetic and evolutionary hypotheses of Zona (1990) based on parsimony analysis of morphological characters. We describe how exon baits designed from 176 putatively single-copy nuclear loci and nearly 90 kb of the chloroplast were used to enrich for these phylogenetically informative targets and generate sequence data to estimate a species-level phylogeny for *Sabal*. The exon baits were designed from several genera from across the Arecaceae and will be widely useful for phylogenomic analyses among all palm lineages.

In addition to describing our methods for bait design and targeted gene capture, we compare methods for multi-locus phylogenetic inference of species-level relationships. As has been widely discussed in the phylogenetics literature (e.g. Kuo, Wares & Kissinger, 2008; Cranston *et al.*, 2009; Salichos & Rokas, 2013), discordance among gene histories is expected due to several biological processes, including hybridization, lateral gene transfer, gene duplication and loss, and incomplete lineage sorting (ILS) (Maddison, 1997; Rosenberg, 2002). These factors can lead to gene topologies with well-supported discordance.

The use of the coalescent model in phylogenetics has allowed discordant gene trees to be informative even with ILS (Maddison & Knowles, 2006). ILS is commonly observed through comparisons of gene tree and species tree topologies (e.g. Takahashi *et al.*, 2001; Morando *et al.*, 2004; Pollard *et al.*, 2006), especially when internal branches of a species tree are short (e.g. rapid radiations) and genetic drift or selection has not had sufficient time to remove ancestral polymorphism (Maddison, 1997). Species tree methods based on the multi-species coalescent aim to maximize the probability of a distribution of gene trees given a species tree, recognizing that gene tree ancestries may coalesce at different times but that all genes must coalesce earlier than species (Degnan & Rosenberg, 2009). Approaches vary in their implementation: many estimate species relationships using the coalescent process and can account for gene tree uncertainty (Liu, 2008; Kubatko, Carstens & Knowles, 2009; Heled & Drummond, 2010; Liu, Yu & Edwards, 2010) while others are based on statistical summaries of variation in gene trees (Baum, 2007; Liu *et al.*, 2009; Larget *et al.*, 2010). Here we investigate how phylogenetic inference based on coalescent species tree estimation procedures compares with those based on analyses of concatenated sequence alignments.

METHODS

Tissue sampling

Leaf tissue samples from all 15 species of *Sabal* as well as two outgroup taxa were obtained from three living collections; *Sabal mexicana* and *S. yapa* were collected from Fairchild Tropical Botanic Garden (Coral Gables, FL, USA), *S. pumos* and *S. rosei* from Nong Nooch Tropical Garden (Pattaya, Thailand), and the remaining 11 *Sabal* species plus the two outgroup species [*Bactris major* and *Dictyosperma album* (subfamily Arecoideae)] from Montgomery Botanical Garden (Coral Gables, FL, USA). Voucher information for all samples is found in Table 1.

Sabal domingensis plastome sequencing, assembly and annotation

DNA was isolated from silica-dried tissue of a single accession of *Sabal domingensis* from Fairchild Tropical Botanical Garden (Voucher [FTBG] 95371B) to create a reference plastome sequence for *Sabal*. One microgram of high-molecular-weight, total genomic DNA was sent to Cold Spring Harbor Laboratory for library preparation, barcoding and multiplex sequencing of 96-bp reads on a single lane of an Illumina GAIIx sequencer (including other barcoded samples). Adapter and barcode trimming were completed at Cold Spring Harbor Laboratory. Reads were then quality trimmed from their 3′ ends using Trimmomatic v. 0.17 (Lohse *et al.*, 2012) with a minimum Phred score of 20 and a sliding window of 4 bases.

Table 1. Specimen voucher information

Species	Living collection accession ID	Voucher ID	Herbarium
Sabal bermudana	MBC 20090159	Larry Noblick 5633	FTG*
Sabal causiarum	MBC 20030267 D	Larry Noblick 5588	FTG
Sabal domingensis	MBC 20120482	Brett Jestrow 2012-207	FTG
Sabal etonia	MBC 948 D	Larry Noblick 5586	FTG
Sabal guatamalensis	MBC 941048.E	Larry Noblick 5630	FTG
Sabal maritima	MBC 931029 G	Larry Noblick 5584	FTG
Sabal mauritiiformis	MBC 94572 P	Larry Noblick 5585	FTG
Sabal mexicana	FTG 59495 B	Larry Noblick 5634	FTG
Sabal miamiensis	MBC 86513	Nancy Hammer & K. Fanning 116	FTG
Sabal minor	MBC 20050710	Noblick & Bernstein 5452	FTG
Sabal palmetto	MBC 20020918 B	Larry Noblick 5631	FTG
Sabal pumos	BKF 194195	JRC307	BKF†
Sabal rosei	BKF 194196	JRC308	BKF
Sabal uresana	MBC 73395 C	Bogler 1310	FTG
Sabal yapa	FTG 822A	Larry Noblick 5635	FTG
Bactris major	MBC 22070264	FTG Noblick 5467	GA‡
Dictyosperma album	MBC 951241	FTG Noblick 5069	GA

*Fairchild Tropical Botanic Garden (Coral Gables, FL, USA).
†Bangkok Forest Herbarium (Thailand).
‡University of Georgia Herbarium (Athens, GA, USA).

Reads were assembled *de novo* using Velvet v. 1.2.03 (Zerbino & Birney, 2008) using an optimized hash length of 51. Minimum contig length was set to 200 bp, while low coverage cutoff was set to 5× 'kmer coverage', which corresponds to ~10× base coverage (see Zerbino & Birney, 2008). Reference-guided assembly was completed with YASRA v. 2.3 (Ratan, 2009), using *Phoenix dactylifera* (Coryphoideae, GenBank accession no. NC_013991) and *Chamaedorea seifrizii* (Arecoideae, GenBank accession no. JX088667) as reference plastomes.

Contigs from both reference-guided assemblies were merged with *de novo* contigs from Velvet in Sequencher v. 5.2 (Gene Codes). Any discrepancies between reference-guided and *de novo* assemblies were identified in Sequencher and corrected. Discrepancies were due to rare (<5%), presumably erroneous base calls in reads that were included in the reference-based *de novo* assembly. The corrected draft plastome was used in a second round of YASRA assembly as a reference for the original reads, and the resulting contigs were reassembled along with *de novo* contigs as before to generate a completed plastome. Reads were mapped back to the completed plastome using the Burroughs–Wheeler Aligner (Li & Durbin, 2009) to get an accurate estimate read depth in the single-copy and inverted repeat (IR) regions. IR boundaries were hypothesized based on sharp changes in read coverage and were verified in Sequencher by aligning reads straddling all four boundaries.

All intron-containing gene sequences and tRNAs were extracted from *Phoenix dactylifera* (GenBank accession no. GU811709) and aligned to the *Sabal domingensis* plastome to serve as a guide for annotation of intron/exon and tRNA boundaries. Annotation of the completed plastome was undertaken in DOGMA (http://dogma.ccbb.utexas.edu/; Wyman, Jansen & Boore, 2004), and a feature table was exported from DOGMA for further annotation and validation in SEQUIN (http://www.ncbi.nlm.nih.gov).

RNA BAIT DESIGN

To capture more than 100 genes simultaneously without resorting to whole genome shotgun sequencing, we designed biotinylated RNA baits complementary to exons of putatively single-copy nuclear genes and the large single-copy region of the *S. domingenesis* plastid genome. Chloroplast probes were designed across the large single-copy region and a portion of the IR, beginning at position 1 and ending at position 101 699 (Fig. 1). To identify single-copy nuclear genes, protein-coding sequences were compiled from five species representing three of the five palm subfamilies: *Sabal bermudana* (Coryphoideae), *Phoenix*

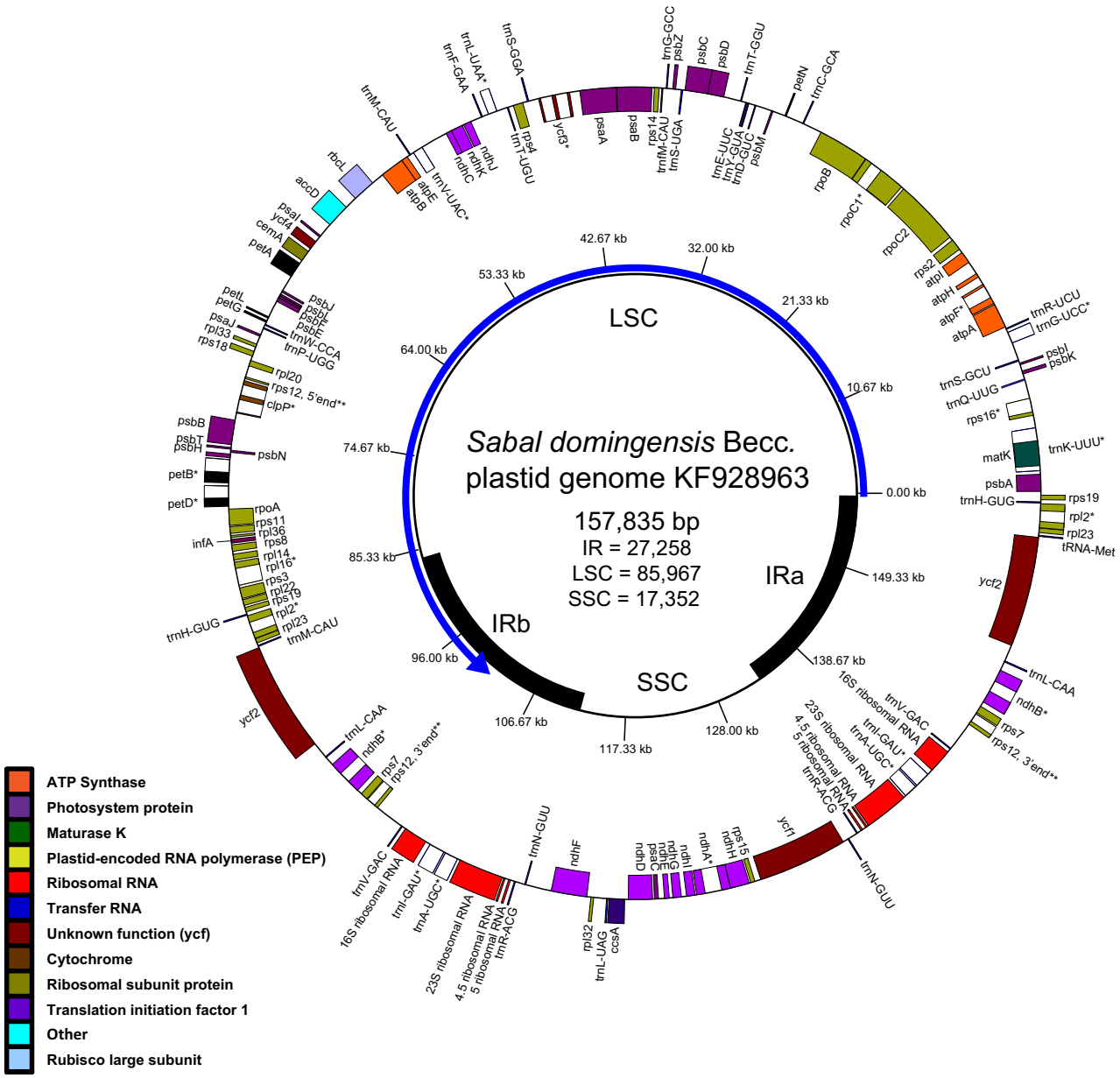

Figure 1. Annotated chloroplast genome of *Sabal domingensis*. The blue arrow indicates the region used for bait design.

dactylifera (Coryphoideae), *Nypa fruticans* (Nypoideae), *Elaeis guineensis* (Arecoideae) and *Cocos nucifera* (Arecoideae). Transcriptome assemblies for *Sabal*, *Nypa* and *Cocos* were obtained from the OneKP project (http://www.onekp.com; Johnson *et al.*, 2012) and assemblies were downloaded from http://www.biomemb.cnrs.fr/contigs.html for *Elaeis* (Bourgis *et al.*, 2011). The protein-coding sequences annotated in the draft *Phoenix dactylifera* genome sequence were downloaded from the genome project website (http://qatarweill.cornell.edu/research/datepalmGenome/download.html; Al-Dous *et al.*, 2011). Using BLASTx (Altschul *et al.*, 1990), protein coding sequences from all five species were sorted into orthogroups containing sequences from ten land plant genomes (http://fgp.bio.psu.edu/tribedb/10_genomes/; Wall *et al.*, 2008). Putatively single-copy gene families were identified as those with single representatives from each of the ten plant genomes (Duarte *et al.*, 2010). Genes from the five palm data sets sorting into 910 single-copy orthogroups were aligned and the pairwise distance was calculated. Putatively single-copy orthogroups were

eliminated from consideration if there was evidence of duplicates in the palm sequences or if the average uncorrected pairwise distance between any two genera was >0.1. The longest palm sequence in each orthogroup was then used as the template for exon–intron boundary annotation using rice exon boundaries. A total of 837 individual exon sequences from 176 nuclear genes were sent to MYcroarray for design of 120-bp RNA baits spanning each exon with 60-bp overlap between adjacent baits. Probes designed against the *S. domingensis* chloroplast were tiled in the same fashion as the nuclear baits, but synthesized separately.

Library construction and sequencing

DNA samples were extracted using a modified CTAB method (Doyle, 1987). Samples were then sheared with a Covaris sonicator to an average insert size of 350 bp and prepared into Illumina libraries with an in-house protocol (modified from Fisher *et al.*, 2011). The random shearing of starting DNA ensured that adjacent intron sequence would be captured with the targeted exons. Nuclear and chloroplast RNA baits were hybridized in the same reaction as prepared Illumina libraries and selectively captured with streptavidin-coated magnetic beads, following the MYbaits protocol (http://www.mycroarray.com). The protocol was modified so that both nuclear and chloroplast baits were added in the same hybridization reaction. Because only 200 kb of nuclear sequence was targeted while the MYbaits protocol is designed for a 2-Mb target, we used one-tenth the nuclear bait concentration specified by the protocol. Libraries were enriched either singularly or in a pool of four samples. Enrichment of targets was verified by quantitative PCR of four targeted regions (three nuclear, one plastid), which were compared with the abundance of a non-target control [a portion of the ribosomal internal transcribed spacer (ITS)] in both enriched pools and un-enriched libraries.

Sequencing was completed in three batches on an Illumina Miseq (150-bp paired reads), with the exception of *S. domingensis*, which was run with 250-bp paired reads. The first sequencing run consisted of two species (*S. causiarum* and *S. mauritiiformis*) run in triplicate to examine the efficiency of different plastid bait and blocking reagent concentrations. The blocking reagent consists of oligonucleotides complementary to Illumina adapter sequences and is used to minimize off-target hybridization. The results of the first sequencing run determined that a blocker concentration of 100 μM was optimal and this concentration was used on all subsequent libraries. Three ratios of chloroplast/nuclear baits were also compared in the first run (1:10, 1:100 and 1:1000). The concentration of plastid baits in the 1:10 treatment resulted in an over-enrichment of plastid reads relative to nuclear reads. Therefore, the remaining enrichments were done with 1:100 or 1:1000 plastid to nuclear bait ratios.

Assembly

Raw sequence reads were sorted by barcode and trimmed at the 3′ end to remove bases with Phred scores less than 20. Reads shorter than 40 bp after trimming or with Phred scores below 20 for more than 20% of the read length were removed from further consideration. The filtered reads for each species were then assembled using combinations of both *de novo* and referenced-based methods. Chloroplast reads were assembled using YASRA (Ratan, 2009), with the *Sabal domingensis* plastome as a reference and a minimum of 3× coverage. For nuclear data, *de novo* and reference-based assemblies were generated using Trinity (Grabherr *et al.*, 2011) and the Columbus algorithm within Velvet (Zerbino & Birney, 2008), respectively. Exon sequences that were submitted for bait design were used as the reference sequence. Trinity isoforms were assessed using the RSEM pipeline and removed if they had no read support or less than 1% of a component's total reads. The Velvet contigs were assembled using a hash size of 31, as determined by VelvetOptimiser (Gladman & Seemann, 2012), and were further extended into intron sequence using SSAKE (Warren *et al.*, 2007). The Trinity and Columbus–Velvet assemblies for each species were then merged using CAP3 (95% identity over 20 bp) (Huang & Madan, 1999). Only the merged sequences were kept (singleton contigs were ignored).

The BLAST program was used to match merged contigs against the exon reference and contigs were renamed according to their best hit. To simplify phylogenetic inference with the nuclear dataset, if two different contigs from a species had best hits to the same exon, they were both removed from the dataset, as allelic phase or paralogy could not be de-convoluted (although alleles or paralogues with >95% identity may have been merged by CAP3). Because original exons used for RNA bait design were annotated with their sequential exon number, contigs matching those exons could be ordered and concatenated to form a gene scaffold before orthologous genes were aligned across species. The percentage of reads mapped was calculated using Bowtie2 (Langmead & Salzberg, 2012) with the '–local' option and treating reads as unpaired, and coverage statistics were calculated using BEDtools (Quinlan & Hall, 2010).

To decrease the distance between ingroup members and the outgroup species, we included sequence

data from *Phoenix dactylifera* (Al-Dous *et al.*, 2011, GenBank accession no. ACYX00000000). BLASTn matches to the 176 single-copy gene exons and the chloroplast genome sequences were extracted from the *P. dactylifera* scaffold assemblies. *Sabal* chloroplast contigs obtained through sequence capture, and available genome sequences of *P. dactylifera* were aligned in MAFFT v.7 (Katoh & Standley, 2013). Plastid alignments were trimmed to include only the region that had been targeted by our RNA baits. Nuclear genes were aligned in PRANK (Löytynoja & Goldman, 2005) with the reference exon sequences that were used for bait design included as a guide. Nuclear alignments were filtered with Gblocks (Castresana, 2000) to remove non-conserved and poorly aligned regions. Given the gapped nature of the nuclear gene alignments due to variable coverage of intron regions, we excluded genes that had an average pairwise distance of greater than 0.4. The final set of 133 nuclear genes used had low frequencies of missing data, with only two genes having less than two-thirds of taxa present.

Phylogenetic analysis

Gene trees were separately estimated for the plastome and the 133 nuclear genes using RAxML's rapid bootstrapping algorithm (Stamatakis, 2006; Stamatakis, Hoover & Rougemont, 2008) with 500 bootstrap replicates per gene and the GTRGAMMA model of nucleotide substitution. The appropriate model of evolution was determined by running JModelTest [using Akaike information criterion (AIC) (Darriba *et al.*, 2012) on four randomly selected nuclear gene alignments from the 133 genes and the entire plastome alignment. We used STAR (Liu *et al.*, 2009) and MP-EST (Liu *et al.*, 2010) to estimate species trees from the nuclear gene trees, with bootstrap tree re-sampling invoked by each program. STAR was implemented on the STRAW server (http://bioinformatics.publichealth.uga.edu/SpeciesTreeAnalysis/index.php; Shaw *et al.*, 2013) and uses average ranks of coalescence times of taxa to construct a minimal distance tree. MP-EST was implemented locally and uses a pseudo-likelihood estimator to generate a species tree. Nuclear gene alignments were also concatenated and the resulting supermatrix was analysed using the GTRGAMMA model in RAxML.

*BEAST 1.7.5 (Heled & Drummond, 2010) was used to concurrently estimate gene trees with the species tree. Because of the computational requirements of *BEAST, we binned genes randomly into groups of five, making concatenated alignments of 'supergenes'. Previous work has shown that limited binning can approximate analyses where each gene is treated independently, even under ILS (Bayzid & Warnow, 2013; Zimmermann, Mirarab & Warnow, 2014). Binning increases the phylogenetic signal for each set of five concatenated gene alignments, resulting in better phylogenetic estimation (Salichos & Rokas, 2013). In BEAUTi 1.7.5, each concatenated alignment was treated as a separate supergene with independent trees (Yule process), substitution rates (relaxed clock) and HKY model parameters. Analysis of each concatenated alignment was seeded by a UPGMA starting tree. Two independent runs were conducted, each to 500 million generations. Convergence was assessed by Tracer v1.5 (Rambaut & Drummond, 2009b) and the two runs were combined after a burnin of 75 million of each run for a combined total of 850 milion states. Effective sample sizes (ESS) of parameters, which indicate the size and therefore the accuracy of the posterior distribution for that parameter, were checked for values > 200, although the ESS of parameters not of immediate interest were allowed to be slightly lower (i.e. population size parameters). The final maximum clade credibility tree was calculated from the combined posterior distribution using TreeAnnotator (Rambaut & Drummond, 2009a).

RESULTS

Sequencing the *Sabal domingensis* plastome yielded a mean coverage of 29.6× across the final plastome assembly of 157 835 bp. The plastome had the typical angiosperm chloroplast genome structure, including an IR region (27 258 bp) separated by large and small single-copy regions (85 967 and 17 352 bp, respectively, Fig. 1, GenBank accession no. KF928963).

Sequencing target-enriched libraries in three batches produced high variability of read counts and coverage between libraries (Table 2). Across all species, an average of 159 of the 176 targeted nuclear genes were successfully captured and assembled, with total length of assembled, captured sequence (both exon and intron) ranging from 48 965 to 355 729 bp. Number of reads (pre- and post-filtering), number of contigs after CAP3, number of contigs matching exon sequences, percent on-target reads, coverage, and number of genes per library are reported in Table 2. Concentrations of plastid probes were also varied for libraries enriched with 1:100 or 1:1000 ratios of plastid to nuclear baits. Libraries enriched with a 1:1000 ratio had an average of 1.59% reads mapped to the chloroplast targeted assembly and 29.83× coverage, whereas libraries enriched with the 1:100 bait ratio had an average of 15.85% chloroplast reads mapped and 263.18× coverage (Table 3).

Table 2. Sequencing summary information

Library	Raw reads	Filtered*	Contigs†	Exon‡	Mapped (%)§	Genes¶	Cov**	Cov††
Sabal causiarum	2176 616	1148 083	1615	422	8.47	166	493.5	61.68
Sabal mauritiiformis	1578 616	829 592	1329	413	10.78	164	63.4	13.9
Sabal bermudana	1433 840	1316 243	2618	401	21.37	163	13.28	4.32
Sabal rosei	3059 760	2780 108	26 763	464	52.51	165	201.02	47.96
Sabal maritima	337 558	308 496	573	209	21.21	124	7.17	3
Sabal mexicana	1497 844	1274 719	18 991	449	66.82	166	90.3	30.67
Sabal etonia	2618 138	2387 557	26 155	498	69.7	166	168.43	47.28
Sabal guatamalensis	2153 702	1829 167	24 755	497	63.99	167	133.16	33.96
Sabal palmetto	713 656	614 978	13 299	462	48.55	164	33.69	9.01
Sabal yapa	789 776	724 734	22 659	477	74.81	167	34.38	10.08
Sabal minor	1412 612	1347 161	8827	439	40.56	166	18.08	5.62
Sabal miamensis	868 218	835 033	2004	411	47.59	167	17.09	4.99
Sabal domingensis	3528 060	2799 679	6308	377	27.96	162	27.86	10.17
Sabal pumos	1381 994	1289 257	4625	122	36.52	92	6.21	3.13
Sabal uresana	1380 898	1309 349	12 204	438	67.69	164	20.66	5.93
Dictyosperma album	494 262	437 338	12 791	461	44.21	168	21	6.36
Bactris major	2626 098	2473 322	61 848	445	39.98	169	88.61	19.03

*Read counts after cleaning.
†Total number of assembled contigs merged by CAP3.
‡Number of assembled targeted exons.
§Percentage of reads which map to exons.
¶Number of total targeted genes assembled.
**Coverage across exon regions.
††Coverage across intron regions.

Table 3. Plastid assembly and coverage of targeted and non-targeted portions

			Targeted region		Non-target plastome	
Library	Plastid conc.	Contigs	Reads mapped (%)	Cov.	Reads mapped (%)	Cov.
Sabal causiarum	1:1000	28	3.16	53.97	1.36	36.90
Sabal mauritiiformis	1:100	61	2.07	24.83	1.06	20.27
Sabal bermudana	1:100	30	26.97	541.97	8.71	282.31
Sabal rosei	1:1000	44	2.32	95.58	1.22	80.92
Sabal maritima	1:100	22	10.70	46.60	3.73	26.31
Sabal mexicana	1:1000	61	0.56	9.14	0.35	186.57
Sabal etonia	1:1000	112	0.61	19.20	0.35	19.94
Sabal guatamalensis	1:1000	114	0.71	17.65	0.40	15.88
Sabal palmetto	1:1000	83	1.31	11.66	0.79	11.14
Sabal yapa	1:1000	106	1.39	12.90	0.88	13.02
Sabal minor	1:100	39	24.23	495.33	8.25	273.06
Sabal miamiensis	1:100	13	20.11	234.02	6.93	131.20
Sabal domingensis	1:100	7	54.77	3068.23	20.62	1885.98
Sabal pumos	1:100	23	11.19	215.88	4.63	143.02
Sabal uresana	1:1000	18	15.71	22.35	6.28	186.57
Dictyosperma album	1:1000	53	1.33	48.42	0.79	19.02
Bactris major	1:100	80	15.71	283.66	1.89	46.20

Some non-targeted regions of the plastome could be assembled but, as expected, coverage was lower than observed for the targeted region (Table 3).

The cleaned chloroplast alignment (87 400 bp) had a maximum uncorrected pairwise distance of 0.016 (average of 0.0048) between sequences and 598

(0.68%) parsimony-informative sites, of which 334 fell into gene coding regions, which accounted for 60 200 of the 87 400-bp plastome alignment. Of the 598 total parsimony-informative sites, just 126 (21%) were found between ingroup species (*Sabal* only). The plastome-derived gene tree places *S. yapa* and *S. mauritiiformis* as sister to all other *Sabal* species with high support (Fig. 2B), but is otherwise unable to resolve relationships among *Sabal* species.

Species trees derived from coalescent-based analyses of nuclear genes using STAR, MP-EST and *BEAST (Fig. 3A) were largely consistent with one another. The branch lengths shown for the *BEAST tree were not calibrated due to the lack of unambiguous New World fossils for *Sabal*. Nonetheless, internal branch lengths in the uncalibrated *BEAST chronogram (Fig. 3B) are short, indicating rapid diversification while explaining gene tree discordance (Fig. 4). Gene trees discordant with particular nodes in the estimated species trees (Fig. 3) do not share a common alternative resolution, suggesting random loss and retention of ancestral variation as modelled by the multispecies coalescent. Interestingly, comparison of all the topologies shown in Figure 3 indicates that the supermatrix analysis – which assumes a common history for all concatenated genes – recovered a tree (Fig. 3C) that is largely consistent with those from coalescent analyses. This is not too surprising, given the weak phylogenetic signal in most of the gene trees.

Digging more deeply into the nature of incongruence among gene trees, rooted gene trees were queried for the resolution of relationships and level of support for specific clades observed in the species trees (Figs 3, 4). Support and conflict are defined as follows: a gene tree strongly supported the species clade if the queried node in the gene tree had ≥80% bootstrap support; the gene tree weakly supported the queried node if the node had ≥50% but <80% bootstrap support; gene trees that had node support of >20% and <50% were considered to be weak conflict; gene trees with ≤20% support for the queried node were classified as strong conflict. Nodes with missing data in a given gene tree were included in these counts so long as more than one taxon from the queried node was present. Gene tree nodes where one taxon or fewer from the queried species was present were categorized as 'missing'. All tree queries were performed using custom scripts, which are deposited on GitHub (https://github.com/kheyduk). While deeper nodes had few incongruent genes (i.e. the entire *Sabal* clade, Fig. 4), more recent nodes had high levels of gene tree incongruence.

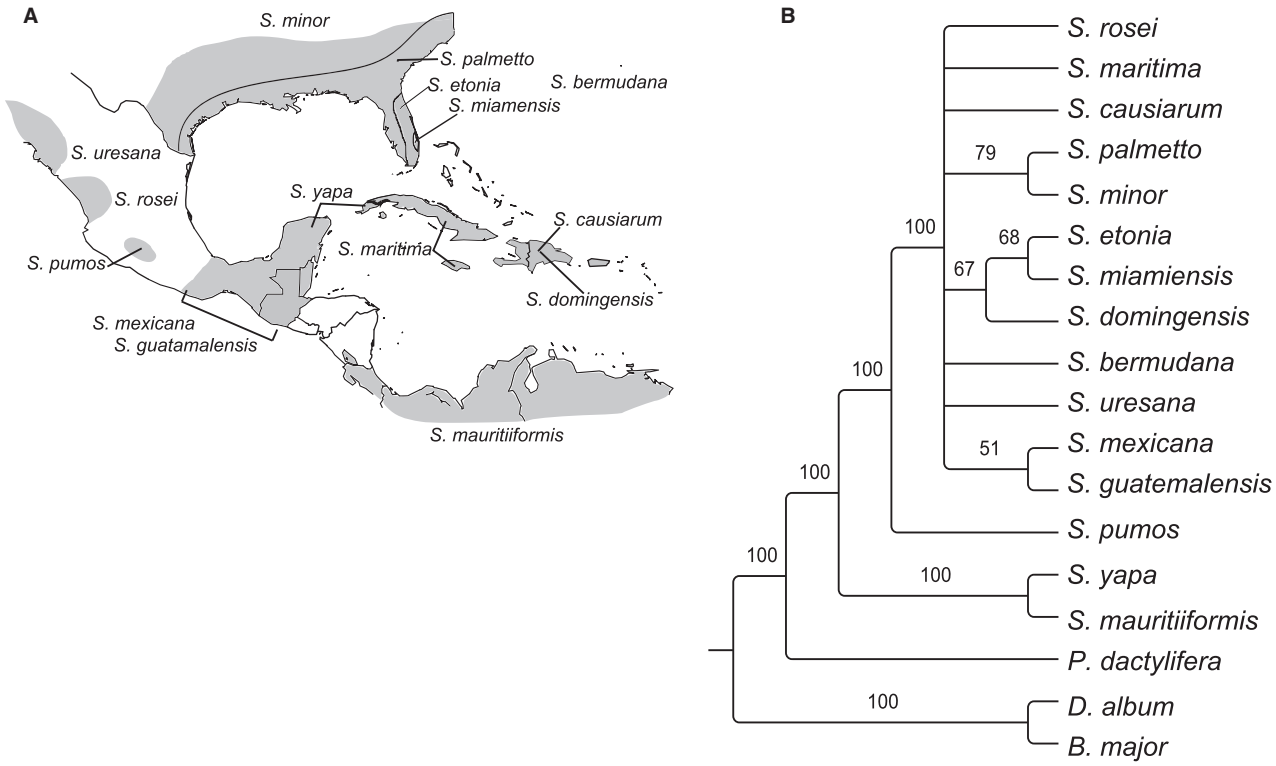

Figure 2. A, distribution of the species in *Sabal*. Ranges based on Zona (1990) and Henderson, Galeano & Bernal (1995); B, plastome maximum-likelihood tree estimate based on an aligned region of 87 400 bp.

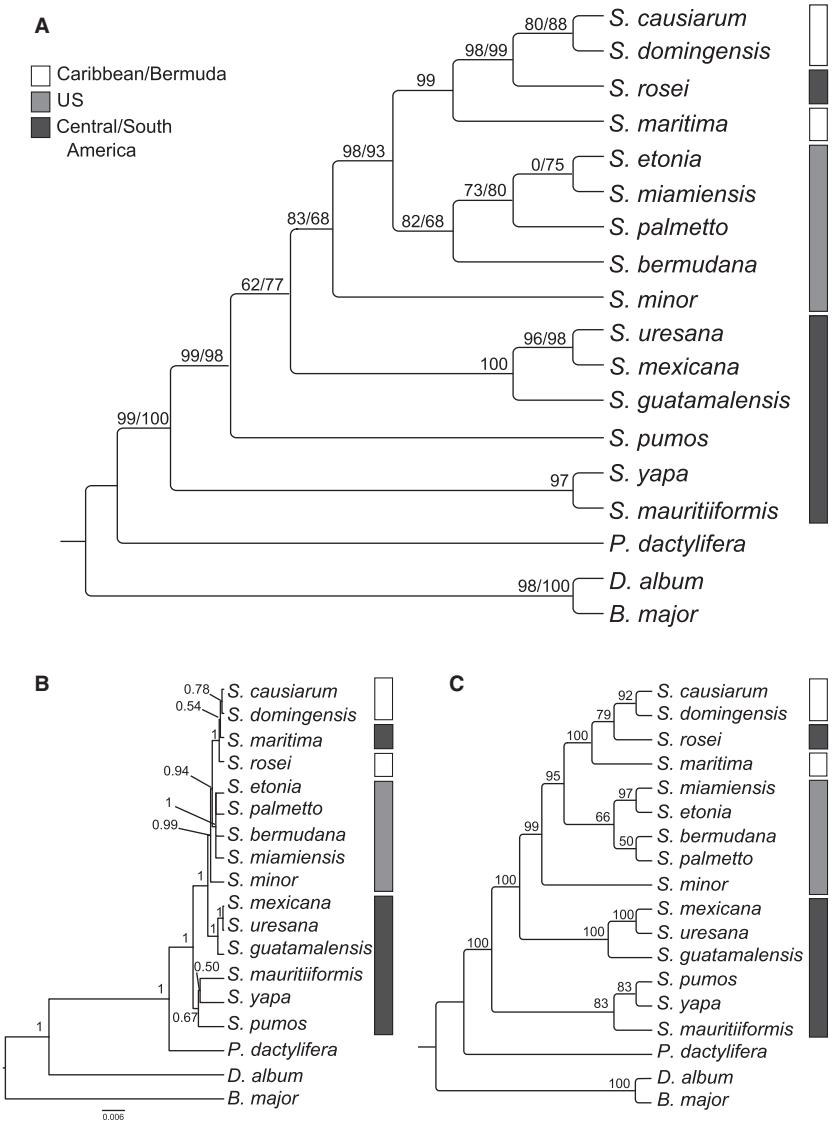

Figure 3. A, STAR and MP-EST trees share largely identical topologies but differ in support. The first support values are from MP-EST and the second from STAR (if only one value, both estimation methods had the same support); B, *BEAST phylogram with posterior probabilities – nodes with posterior probability < 50 were collapsed; C, concatenated dataset maximum-likelihood tree estimate.

DISCUSSION

Multi-locus species tree estimation

The *Sabal* plastome sequences have low divergence and the tree estimated from the plastome sequences shows low resolution and low support values (Fig. 2). Whole plastid genomes have been useful for assessing relationships among flowering plant orders and families (e.g. Jansen *et al.*, 2007; Givnish *et al.*, 2010) and to a lesser degree within genera (Parks, Cronn & Liston, 2009). The Arecaceae have been noted for low substitution rates (Wilson, Gaut & Clegg, 1990; Gaut *et al.*, 1996) and our results illustrate the limited utility of the conserved plastome for resolving relationships within palm genera. Furthermore, while sequencing large portions of the plastome offers many kilobases for analysis of structural rearrangements and selective signals in addition to nucleotide substitutions, it is a single locus and random sorting of ancestral variation may support a plastome history that does not reflect the speciation history within the genus. The inability of over 90 kb of plastid sequence to resolve relationships within *Sabal* underlines the need for a multilocus nuclear approach for resolution of palm species relationships.

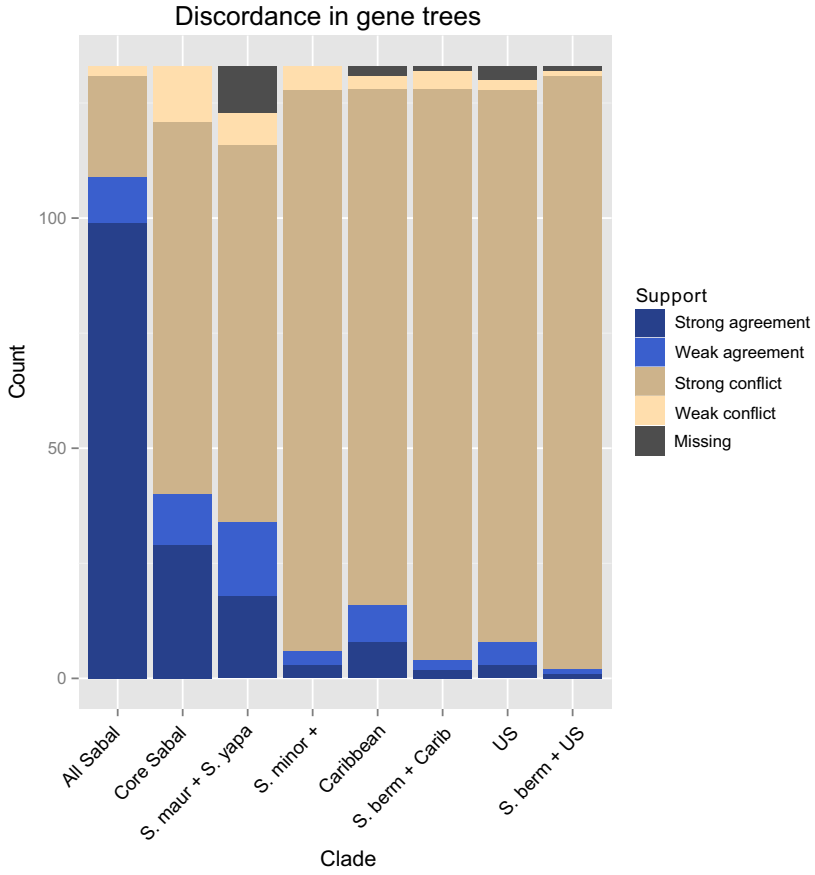

Figure 4. Bar graph representing the number of gene trees that strongly support or reject a clade. Each column represents a specific clade: *Sabal* (all *Sabal* species), Core *Sabal* (*Sabal* species excluding *S. mauritiiformis* and *S. yapa*), *S. maur* + *S. yapa* (clade which includes *S. mauritiiformis* and *S. yapa*), *S. minor* + (all taxa above and including *S. minor*), Caribbean (*S. causiarum*, *S. domingensis*, *S. rosei*, *S. maritima*), *S. berm* + Carib (*S. bermudana* sister to *S. rosei*, *S. maritima*, *S. domiigensis* and *S. causiarum*), USA clade (*S. etonia*, *S. miamensis* and *S. palmetto*) and *S. berm* + USA (*S. bermudana* sister to *S. etonia*, *S. miamensis* and *S. palmetto*).

The three methods used to estimate species relationships from nuclear gene trees – STAR, MP-EST and *BEAST – have topologies that are highly congruent despite being fundamentally different in their approaches to species tree estimations. STAR estimates a species tree from a distance matrix of average ranks of nodes. The average rank of coalescence is shown to approach the expected rank of coalescence as the number of genes increases, and is therefore a good estimate of the species tree in multi-locus datasets (Liu *et al.*, 2009). In contrast, MP-EST attempts to maximize the pseudo-likelihood of a species tree given a set of gene trees. Pseudo-likelihood of a species tree is estimated by the frequencies of rooted triples in the gene trees (Liu *et al.*, 2010). *BEAST simultaneously estimates gene trees and the species tree from sequence data, and additionally can estimate branch lengths in terms of effective population size if multiple samples per taxon are used. We were limited to just one sample per species, so branch lengths shown in the *BEAST tree represent an uncalibrated degree of divergence between clades.

Although the algorithms for species tree estimation in STAR and MP-EST differ, their consistency in species tree estimation is perhaps unsurprising given that both programs were run on the same set of bootstrapped gene trees (Fig. 3A). *BEAST, by contrast, recalculated gene trees from the initial alignments, and therefore is an independent estimation of gene trees and the species tree (Fig. 3B). MP-EST was unable to resolve the *S. etonia* and *S. miamiensis* split resolved by STAR (indicated by 0 bootstrap support in Fig. 3A). *BEAST differed in its placement of *S. pumos*, grouping it sister to *S. yapa* and *S. mauritiiformis*. The concatenated phylogeny is largely congruent with the STAR, MP-EST and *BEAST estimates, and places *S. pumos* sister to

S. yapa and *S. mauritiiformis* as with *BEAST (Fig. 3C). Recent work on multi-locus data has shown that concatenation can converge with high confidence on the wrong topology, particularly as the density of short internodes increases, due either to rapid radiations or to large numbers of taxa (Degnan & Rosenberg, 2006, 2009; Kubatko & Degnan, 2007; Huang & Knowles, 2009; Rosenberg, 2013). The moderately supported conflict among species trees in the placement of *S. pumos* relative to *S. yapa*, *S. mauritiiformis* and a clade with the remaining *Sabal* species may be an artefact of concatenation of all genes in the supermatrix analysis (Fig. 3C) and random sets of five genes in the *BEAST analysis (Fig. 3B) vs. no concatenation of gene alignments in the MP-EST and STAR analyses (Fig. 3A). We interpret this conflict with care pending improved understanding of the impact of binning gene alignments for species tree estimation (e.g. Zimmermann *et al.*, 2014).

Relatively few gene trees agree with both the coalescent and the concatenation methods, highlighting how robust these analyses are to ILS under some conditions (Fig. 4). Nodes deeper in the tree (i.e. the *Sabal* clade) are supported by many gene trees, but topological support among gene trees decreases as node ages decrease. Incongruence among gene trees is perhaps unsurprising given the short branch lengths in the *BEAST tree (Fig. 3B). Maddison & Knowles (2006) showed that sampling multiple individuals can mitigate problems of deep coalescence and the resulting gene tree incongruence, particularly in shallow species trees. They note, however, that the benefit of more individuals per species appears to plateau as the number of loci increases. Here, the use of a single individual per taxon does not appear to hinder our ability to estimate the species tree of *Sabal*. Sampling more individuals per species could, however, further improve support for relationships within *Sabal* and provide insight into the size of extant and ancestral populations. Moreover, increasing the number of samples per species in future analyses of *Sabal* and other palms could help to differentiate whether the observed gene tree discordance is caused by ILS or gene flow. Interspecific hybridization has been implicated within *Sabal*; Goldman *et al.* (2011) characterized *Sabal* × *brazoriensis* D.H. Goldman, L. Lockett, & R.W. Read, nothosp. nov. as a late-generation hybrid, morphologically and genetically distinct from putative parents *S. minor* and *S. palmetto*.

Sabal divergence and biogeography

Bailey (1944) noted that *Sabal palmetto* should be considered sister to all the *Sabal* species due to a lack of morphological specialization (Zona, 1990). The reassessment of relationships within *Sabal* by Zona used outgroups as a basis for character state polarity, and the resulting cladogram separates *S. minor* from the remainder of the genus. The placement of *S. yapa* and *S. mauritiiformis* as sister to the rest of *Sabal* in the MP-EST/STAR (Fig. 3A) trees is consistent with their occurrence in the current centre of diversity of the genus. A core group of Central American species (*S. mexicana*, *S. uresana*, *S. guatamalensis*, *S. pumos*) is paraphyletic to the USA, Bermuda and Caribbean species in this analysis. The nuclear coalescent and supermatrix trees place *S. minor* as sister to non-Mexican species, indicating the radiation probably went northward into the southern United States, spreading across the Gulf region and southward into the Caribbean islands.

While the Caribbean island species arose as a result of vicariance or short-distance dispersal, *Sabal* colonization of Bermuda probably resulted from long-distance dispersal. The islands of Bermuda exist as an archipelago on the southern margin of a 650-km^2, mostly submerged limestone platform on top of the remnants of a volcanic seamount rising 4000 m from the Atlantic Ocean floor. The volcano was last active in the Oligocene (Reynolds & Aumento, 1974) and the exposed limestone cap is estimated to be 2 Myr old at most (Vacher, Hearty & Rowe, 1995). With a current maximum elevation of ~76 m above sea level, Bermuda's sediment deposits reveal considerable fluctuations in sea level. The highest sea level of 20 m above current levels occurred 400 000 years ago, essentially reducing the islands to 1/20th of their current land area of ~53.2 km^2 (Olson, 2008). At that time, Bermuda may have consisted of a small number of islets, many of which probably experienced storm overwash (Olson & Hearty, 2003), particularly when severe hurricanes struck. It has been suggested that island endemics that could not withstand these conditions are either not truly endemic, or must have arrived in Bermuda within the last 0.5 Myr once ocean waters had subsided and the island stabilized (Olson, Hearty & Pregill, 2006; Olson, 2008). However, the history of subsidence or portions of Bermuda remains controversial. For example, a volcanic seamount may have remained above sea level long after dunes dating to the Pleistocene began to develop, thus creating a temporal land bridge between Tertiary and Pleistocene Bermuda (David B. Wingate, pers. comm.).

Fossil evidence suggests *S. bermudana* dispersal to Bermuda long before the sea-level maximum 400 000 years ago. Fossil leaf impressions have been found in all the high sea stand aeolianites dating to 800 000 to 2 Myr ago (David B. Wingate, pers. comm.) indicating that either *S. bermudana* was able

to survive the highest sea levels or Bermuda was larger than previously thought. That *S. bermudana* is able to tolerate saltwater from storm overwash is supported by the occurrence of extant individuals near shorelines and on islets <1 ha in area (D.W.T., pers. observ.). The early arrival of terrestrial organisms prior to the seal-level maximum is supported by fossils of the endemic skink (*Plestiodon longirorstris*), which indicate the species arrived on the islands 400 000 to 2 Myr ago (Brandley *et al.*, 2010).

The identity of the immediate sister species of *S. bermudana* is not clear. The coalescence trees and the supermatrix analysis place *S. bermudana* within the USA clade with high support (Fig. 3). However, relatively few gene trees support *S. bermudana* with the USA clade or the Caribbean clade. The mixed signal regarding *S. bermudana*'s phylogenetic placement is probably the result of the relatively quick and simultaneous speciation events that occurred in the majority of extant taxa.

Sequence capture

The variability between sequenced libraries in terms of total number of reads and capture efficiency highlights the flexibility of the sequence capture approach. While two libraries (*S. pumos*, *S. maritima*) were particularly poor, they nevertheless were useful in subsequent analyses. Similarly, low percentages of mapped reads and coverage in the nuclear dataset did not always correlate to low mapping and coverage for the plastome assembly; *S. pumos*, one of the lowest quality assemblies for the nuclear baits, had over 100 kb of chloroplast sequence and represents the longest plastid assembly of all the species sequenced. The two outgroup species, *D. album* and *B. major*, are members of the Arecaceae but in the subfamily Arecoideae, which is estimated to have diverged from the Coryphoideae 85 Mya (Baker & Couvreur, 2013). Nevertheless, plastid and nuclear DNA for both species was successfully captured with the RNA baits with efficiencies comparable to members of *Sabal* (Table 2). By designing the baits in conserved exon sequences in an overlapping fashion, we have maximized the utility of the bait set for use across anciently diverged species within Arecaceae. Our probe designs can be used across the family for more detailed biogeographical studies and resolution of relationships among lineages exhibiting little genetic divergence.

This work adds to the growing literature illustrating the utility of targeted sequence capture for resolving relationships among species (Faircloth *et al.*, 2012; Grover *et al.*, 2012; Lemmon *et al.*, 2012; McCormack *et al.*, 2013; Weitemier *et al.*, 2014), even within groups with low nucleotide substitution rates and rapid diversification rates. Well-resolved and robust phylogenetic inferences shed light on the historical biogeography and diversification of *Sabal*. At the same time, even with over 100 nuclear loci and nearly 90 kb of plastid genome sequence, ambiguities remain. Future within-species sampling may elucidate the evolutionary processes responsible for this ambiguity.

ACKNOWLEDGEMENTS

This work was supported by the University of Georgia Office of the Vice President for Research (D.W.T.); UGA Faculty Research Grant (D.W.T.); and a National Science Foundation grant to J.L.-M. (grant number DEB 0829868).

We are especially grateful to Dr Larry Noblick at Montogmery Botanical Garden Center for his invaluable assistance in obtaining samples and making voucher specimens. We also thank Jason Comer, Charlotte Carrigan and Saravanaraj Ayyampalayam for assistance with collection, laboratory work and bioinformatics, respectively, and the reviewers who helped improve the manuscript.

REFERENCES

Al-Dous EK, George B, Al-Mahmoud ME, Al-Jaber MY, Wang H, Salameh YM, Al-Azwani EK, Chaluvadi S, Pontaroli AC, DeBarry J, Arondel V, Ohlrogge J, Saie IJ, Suliman-Elmeer KM, Bennetzen JL, Kruegger RR, Malek JA. 2011. De novo genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). *Nature Biotechnology* **29:** 521–527.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215:** 403–410.

Bacon CD, Baker WJ, Simmons MP. 2012. Miocene dispersal drives island radiations in the palm tribe Trachycarpeae (Arecaceae). *Systematic Biology* **61:** 426–442.

Bailey LH. 1944. Revision of the palmettoes. *Genes Herbarium* **6:** 365–459.

Baker WJ, Couvreur TLP. 2013. Global biogeography and diversification of palms sheds light on the evolution of tropical lineages. I. Historical biogeography. *Journal of Biogeography* **40:** 274–285.

Baker WJ, Dransfield J, Hedderson TA. 2000. Phylogeny, character evolution, and a new classification of the calamoid palms. *Systematic Botany* **25:** 297.

Baker WJ, Savolainen V, Asmussen-Lange CB, Chase MW, Dransfield J, Forest F, Harley MM, Uhl NW, Wilkinson M. 2009. Complete generic-level phylogenetic analyses of palms (Arecaceae) with comparisons of supertree and supermatrix approaches. *Systematic Biology* **58:** 240–256.

Baum DA. 2007. Concordance trees, concordance factors, and the exploration of reticulate genealogy. *Taxon* **56:** 417–426.

Bayzid MS, Warnow T. 2013. Naïve binning improves phylogenomic analyses. *Bioinformatics* **29:** 2277–2284.

Bourgis F, Kilaru A, Cao X, Ngando-Ebongue GF, Drira N, Ohlrogge JB, Arondel V. 2011. Comparative transcriptome and metabolite analysis of oil palm and date palm mesocarp that differ dramatically in carbon partitioning. *Proceedings of the National Academy of Sciences of the United States of America* **108:** 12527–12532.

Brandley MC, Wang Y, Guo X, Montes N, de Oca A, Feria Ortiz M, Hikida T, Ota H. 2010. Bermuda as an evolutionary life raft for an ancient lineage of endangered lizards. *PLoS ONE* **5:** e11375.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* **17:** 540–552.

Cranston KA, Hurwitz B, Ware D, Stein L, Wing RA. 2009. Species trees from highly incongruent gene trees in rice. *Systematic Biology* **58:** 489–500.

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* **9:** 772.

Degnan JH, Rosenberg NA. 2006. Discordance of species trees with their most likely gene trees. *PLoS Genetics* **2:** e68.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution* **24:** 332–340.

Doyle J. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* **19:** 11–15.

Dransfield J, Uhl NW, Asmussen CB, Baker WJ, Harley MM, Lewis CE. 2008. *Genera Palmerum* – the evolution and classification of palms. Kew: Royal Botanic Gardens.

Duarte JM, Wall PK, Edger PP, Landherr LL, Ma H, Pires JC, Leebens-Mack J, dePamphilis CW. 2010. Identification of shared single copy nuclear genes in Arabidopsis, Populus, Vitis and Oryza and their phylogenetic utility across various taxonomic levels. *BMC Evolutionary Biology* **10:** 61.

Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology* **61:** 717–726.

Fisher S, Barry A, Abreu J, Minie B, Nolan J, Delorey TM, Young G, Fennell TJ, Allen A, Ambrogio L, Berlin AM, Blumenstiel B, Cibulskis K, Friedrich D, Johnson R, Juhn F, Reilly B, Shammas R, Stalker J, Sykes SM, Thompson J, Walsh J, Zimmer A, Zwirko Z, Gabriel S, Nicol R, Nusbaum C. 2011. A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biology* **12:** R1.

Gaut BS, Morton BR, McCaig BC, Clegg MT. 1996. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. *Proceedings of the National Academy of Sciences of the USA* **93:** 10274–10279.

Givnish TJ, Ames M, McNeal JR, McKain MR, Steele PR, dePamphilis CW, Graham SW, Pires JC, Stevenson DW, Zomlefer WB, Briggs BG, Duvall MR, Moore MJ, Heaney JM, Soltis DE, Soltis PS, Thiele K, Leebens-Mack JH. 2010. Assembling the tree of the monocotyledons: plastome sequence phylogeny and evolution of Poales 1. *Annals of the Missouri Botanical Garden* **97:** 584–616.

Gladman S, Seemann T 2012. VelvetOptimiser. Available at: http://www.vicbioinformatics.com/software.velvetoptimiser.shtml

Goldman DH, Klooster MR, Griffith MP, Fay MF, Chase MW. 2011. A preliminary evaluation of the ancestry of a putative *Sabal* hybrid (Arecaceae: Coryphoideae), and the description of a new nothospecies, *Sabal* × *brazoriensis*. *Phytotaxa* **27:** 8–25.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson D, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29:** 644–652.

Grover CE, Salmon A, Wendel JF. 2012. Targeted sequence capture as a powerful tool for evolutionary analysis. *American Journal of Botany* **99:** 312–319.

Hahn WJ. 2002. A molecular phylogenetic study of the Palmae (Arecaceae) based on *atpB*, *rbcL*, and 18S nrDNA sequences. *Systematic Biology* **51:** 92–112.

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

Henderson A, Galeano G, Bernal R. 1995. *Field guide to the palms of the Americas*. Princeton, NJ: Princeton University Press.

Heyduk K, Trapnell DW, Barrett CF, Leebens-Mack J. 2015. Data from: Phylogenomic analyses of species relationships in the genus *Sabal* (Arecaceae) using targeted sequence capture. *Dryad Digital Repository*. doi: 10.5061/dryad.jm78g.

Huang H, Knowles LL. 2009. What is the danger of the anomaly zone for empirical phylogenetics? *Systematic Biology* **58:** 527–536.

Huang X, Madan A. 1999. CAP3: a DNA sequence assembly program. *Genome Research* **9:** 868–877.

Jansen RK, Cai Z, Raubeson LA, Daniell H, Depamphilis CW, Leebens-Mack J, Müller KF, Guisinger-Bellian M, Haberle RC, Hansen AK, Chumley TW, Lee SB, Peery R, McNeal JR, Kuehl JV, Boore JL. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proceedings of the National Academy of Sciences of the United States of America* **104:** 19369–19374.

Johnson MTJ, Carpenter EJ, Tian Z, Bruskiewich R, Burris JN, Carrigan CT, Chase MW, Clarke ND, Covshoff S, Depamphilis CW, Edger PP, Goh F, Graham S, Greiner S, Hibberd JM, Jordon-Thaden I,

Kutchan TM, Leebens-Mack J, Melkonian M, Miles N, Myburg H, Patterson J, Pires JC, Ralph P, Rolf M, Sage RF, Soltis D, Soltis P, Stevenson D, Stewart CN, Surek B, Thomsen CJM, Villarreal JC, Wu X, Zhang Y, Deyholos MK, Wong GKS. 2012. Evaluating methods for isolating total RNA and predicting the success of sequencing phylogenetically diverse plant transcriptomes. (C Quince, Ed.). *PLoS ONE* **7:** e50226.

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

Kubatko LS, Degnan JH. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology* **56:** 17–24.

Kubatko LS, Carstens BC, Knowles LL. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics (Oxford, England)* **25:** 971–973.

Kuo CH, Wares JP, Kissinger JC. 2008. The Apicomplexan whole-genome phylogeny: an analysis of incongruence among gene trees. *Molecular Biology and Evolution* **25:** 2689–2698.

Langmead B, Salzberg S. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9:** 357–359.

Larget BR, Kotha SK, Dewey CN, Ané C. 2010. BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics (Oxford, England)* **26:** 2910–2911.

Lemmon AR, Emme SA, Lemmon EM. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology* **61:** 727–744.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics (Oxford, UK)* **25:** 1754–1760.

Liu L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics (Oxford, UK)* **24:** 2542–2543.

Liu L, Yu L, Pearl DK, Edwards SV. 2009. Estimating species phylogenies using coalescence times among sequences. *Systematic Biology* **58:** 468–477.

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

Lohse M, Bolger AM, Nagel A, Fernie AR, Lunn JE, Stitt M, Usadel B. 2012. RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research* **40:** W622–W627.

Löytynoja A, Goldman N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National Academy of Sciences of the United States of America* **102:** 10557–10562.

Maddison WP. 1997. Gene trees in species trees. *Systematic Biology* **46:** 523.

Maddison WP, Knowles LL. 2006. Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology* **55:** 21–30.

Mandel JR, Dikow RB, Funk VA, Masalia RR, Staton SE, Kozik A, Michelmore RW, Rieseberg LH, Burke JM. 2014. A target enrichment method for gathering phylogenetic information from hundreds of loci: and example from the Compositae. *Applications in Plant Sciences* **2:** 1300085.

McCormack JE, Harvey MG, Faircloth BC, Crawford NG, Glenn TC, Brumfield RT. 2013. A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. *PLoS ONE* **8:** e54848.

Morando M, Avila LJ, Baker J, Sites JW. 2004. Phlogeny and phylogeography of the *Liolaemus darwinii* complex (Squamata: Liolaemidae): evidence for introgression and incomplete lineage sorting. *Evolution* **58:** 842–859.

Olson SL. 2008. *Pireella cymbifolia* (Pterobryaceae) new to the flora, with comments on sea level and other factors influencing the phytogeography of Bermuda. *Journal of Bryology* **30:** 224–226.

Olson SL, Hearty PJ. 2003. Probable extirpation of a breeding colony of Short-tailed Albatross (*Phoebastria albatrus*) on Bermuda by Pleistocene sea-level rise. *Proceedings of the National Academy of Sciences of the United States of America* **100:** 12825–12829.

Olson SL, Hearty PJ, Pregill GK. 2006. Geological constraints on evolution and survival in endemic reptiles on Bermuda. *Journal of Herpetology* **40:** 394–398.

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

Pollard DA, Iyer VN, Moses AM, Eisen MB. 2006. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. (BF McAllister, ed.). *PLoS Genetics* **2:** e173.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* **26:** 841–842.

Rambaut A, Drummond AJ. 2009a. Treeannotator v1.5. Available at: http://tree.bio.ed.ac.uk/software/

Rambaut A, Drummond AJ. 2009b. Tracer v1.5. Available at: http://tree.bio.ed.ac.uk/software/

Ratan A. 2009. *Assembly algorithms for next-generation sequence data*. PhD Dissertation, The Pennsylvania State University, University Park, PA, USA.

Reynolds PR, Aumento FA. 1974. Deep Drill 1972: potassium–argon dating of the Bermuda drill core. *Canadian Journal of Earth Sciences* **11:** 1269–1273.

Rosenberg NA. 2002. The probability of topological concordance of gene trees and species trees. *Theoretical Population Biology* **61:** 225–247.

Rosenberg NA. 2013. Discordance of species trees with their most likely gene trees: a unifying principle. *Molecular Biology and Evolution* **30:** 2709–2713.

Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* **497:** 327–331.

Shaw T, Ruan Z, Glenn T, Liu L. 2013. STRAW: Species TRee Analysis Web server. *Nucleic Acids Research* **41**(W1)**:** W238–W241.

Smith SA, Donoghue MJ. 2008. Rates of molecular evolution are linked to life history in flowering plants. *Science (New York, N.Y.)* **322:** 86–89.

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

Stamatakis A, Hoover P, Rougemont J. 2008. A rapid bootstrap algorithm for the RAxML Web servers. *Systematic Biology* **57:** 758–771.

Takahashi K, Terai Y, Nishida M, Okada N. 2001. Phylogenetic relationships and ancient incomplete lineage sorting among cichlid fishes in Lake Tanganyika as revealed by analysis of the insertion of retroposons. *Molecular Biology and Evolution* **18:** 2057–2066.

Vacher HL, Hearty PJ, Rowe MP. 1995. *Terrestrial and shallow marine geology of the Bahamas and Bermuda*. Geological Society of America Special Paper 300.

Wall PK, Leebens-Mack J, Müller KF, Field D, Altman NS, dePamphilis CW. 2008. PlantTribes: a gene and gene family resource for comparative genomics in plants. *Nucleic Acids Research* **36:** D970–D976.

Warren RL, Sutton GG, Jones SJM, Holt RA. 2007. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics (Oxford, UK)* **23:** 500–501.

Weitemier K, Straub SCK, Cronn RC, Fishbein M, Schmickl R, McDonnell A, Liston A. 2014. Hyb-Seq: combining target enrichment and genome skimming of plant phylogenomics. *Applications in Plant Sciences* **2:** 1400042.

Wilson MA, Gaut B, Clegg MT. 1990. Chloroplast DNA evolves slowly in the palm family (Arecaceae). *Molecular Biology and Evolution* **7:** 303–314.

Wyman SK, Jansen RK, Boore JL. 2004. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics (Oxford, UK)* **20:** 3252–3255.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* **18:** 821–829.

Zimmermann T, Mirarab S, Warnow T. 2014. BBCA: improving the scalability of *BEAST using random binning. *BMC Genomics* **15**(Suppl 6)**:** S11.

Zona S. 1990. A monograph of *Sabal* (Arecaceae: Coryphoideae). *Aliso* **12:** 583–666.

SHARED DATA

Raw reads are available through the Short Read Archive (BioProjectID PRJNA276701). Data for phylogenetic analyses, including alignments and gene trees, was deposited in the Dryad Digital Repository (Heyduk *et al*., 2015). The bioinformatics pipeline described, nicknamed 'reads2trees', is available on GitHub: https://github.com/kheyduk/reads2trees.