Reticulate evolution helps explain apparent homoplasy in floral biology and pollination in baobabs (Adansonia; Bombacoideae; Malvaceae)

Nisa Karimi
*University of Wisconsin-Madison*

Corinne E. Grover
*Iowa State University, corrinne@iastate.edu*

Joseph P. Gallagher
*Iowa State University*

Jonathan F. Wendel
*Iowa State University, jfw@iastate.edu*

Cécile Ané
*University of Wisconsin-Madison*

See next page for additional authors
Reticulate evolution helps explain apparent homoplasy in floral biology and pollination in baobabs (Adansonia; Bombacoideae; Malvaceae)

Abstract
Baobabs (Adansonia) are a cohesive group of tropical trees with a disjunct distribution in Australia, Madagascar, and continental Africa, and diverse flowers associated with two pollination modes. We used custom targeted sequence capture in conjunction with new and existing phylogenetic comparative methods to explore the evolution of floral traits and pollination systems while allowing for reticulate evolution. Our analyses suggest that relationships in Adansonia are confounded by reticulation, with network inference methods supporting at least one reticulation event. The best supported hypothesis involves introgression between A. rubrostipa and core Longitubae, both of which are hawkmoth pollinated with yellow/red flowers, but there is also some support for introgression between the African lineage and Malagasy Brevitubae, which are both mammal-pollinated with white flowers. New comparative methods for phylogenetic networks were developed that allow maximum-likelihood inference of ancestral states and were applied to study the apparent homoplasy in floral biology and pollination mode seen in Adansonia. This analysis supports a role for introgressive hybridization in morphological evolution even in a clade with highly divergent and geographically widespread species. Our new comparative methods for discrete traits on species networks are implemented in the software PhyloNetworks.

Keywords
Comparative methods, Hyb-Seq, introgression, network inference, population trees, reticulate evolution, species tree inference, targeted sequence capture

Disciplines
Ecology and Evolutionary Biology

Comments
This is a manuscript of an article published as Karimi, Nisa, Corrinne E. Grover, Joseph P. Gallagher, Jonathan F. Wendel, Cécile Ané, and David A. Baum. "Reticulate evolution helps explain apparent homoplasy in floral biology and pollination in baobabs (Adansonia; Bombacoideae; Malvaceae)." Systematic biology (2019). doi: 10.1093/sysbio/syz073. Posted with permission.

Authors
Nisa Karimi, Corrinne E. Grover, Joseph P. Gallagher, Jonathan F. Wendel, Cécile Ané, and David A. Baum
TITLE: Reticulate evolution helps explain apparent homoplasy in floral biology and pollination in baobabs (*Adansonia*; Bombacoideae; Malvaceae)

RUNNING HEAD: Reticulation in the Baobabs

Nisa Karimi¹*, Corrinne E. Grover², Joseph P. Gallagher²,³, Jonathan F. Wendel², Cécile Ané¹,⁴, David A. Baum¹,⁵*

¹Department of Botany, University of Wisconsin - Madison, WI, 53706, United States of America

²Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA, 50011, United States of America

³Biology Department, University of Massachusetts, Amherst, MA 01003, United States of America

⁴Department of Statistics, University of Wisconsin - Madison, WI, 53706, United States of America

⁵Wisconsin Institute for Discovery, 330 N Orchard St, Madison, WI 53715, United States of America

*Email: nkarimi@wisc.edu, dbaum@wisc.edu

ABSTRACT

Baobabs (*Adansonia*) are a cohesive group of tropical trees with a disjunct distribution in Australia, Madagascar, and continental Africa, and diverse flowers associated with two pollination modes. We used custom targeted sequence capture in conjunction with new and existing phylogenetic comparative methods to explore the evolution of floral traits and
ANCIENT INTROGRESSION IN THE BAOBABS

pollination systems while allowing for reticulate evolution. Our analyses suggest that relationships in *Adansonia* are confounded by reticulation, with network inference methods supporting at least one reticulation event. The best supported hypothesis involves introgression between *A. rubrostipa* and core Longitubae, both of which are hawkmoth pollinated with yellow/red flowers, but there is also some support for introgression between the African lineage and Malagasy Brevitubae, which are both mammal-pollinated with white flowers. New comparative methods for phylogenetic networks were developed that allow maximum-likelihood inference of ancestral states and were applied to study the apparent homoplasy in floral biology and pollination mode seen in *Adansonia*. This analysis supports a role for introgressive hybridization in morphological evolution even in a clade with highly divergent and geographically widespread species. Our new comparative methods for discrete traits on species networks are implemented in the software PhyloNetworks.

**Keywords:** Comparative methods, Hyb-Seq, introgression, network inference, population trees, reticulate evolution, species tree inference, targeted sequence capture
It is widely accepted that hybridization and its genetic consequence, introgression, are widespread phenomena in plants (Stebbins 1959; Arnold, 1992; Reiseberg and Wendel, 1993; Soltis & Soltis, 2009; Payseur & Rieseberg, 2016), and equally in other clades, including animals (Mallett, 2008; Schwenk et al., 2008; Payseur & Rieseberg, 2016). Furthermore, it has also long been appreciated that introgression can be an agent of adaptive evolution in cases where beneficial traits are transferred from a donor to a recipient taxon (Arnold 2004; Reiseberg and Wendel, 1993; Arnold and Kunte, 2017; Suarez-Gonzalez et al. 2018). These facts raise the possibility that adaptive introgression could sometimes carry ecologically important traits between species, and that such events might explain apparent homoplasy when these traits are mapped onto binary trees that do not incorporate the full hybridization history. However, we know relatively little about this possibility. Macroevolutionary methods for systematically detecting reticulation using phylogenomic data are relatively new (Yu et al. 2011, 2014; Park and Nakhleh, 2012; Solís-Lemus et al. 2016; Zhang et al. 2017), and still not widely deployed due to the large amounts of data they require and their computational intensity. Furthermore, comparative methods for ancestral state reconstruction on phylogenetic networks are needed to assess the likelihood that a trait was acquired via a minor hybridization edge (i.e, an edge contributing < 50% of the genes to a hybrid descendant), but such methods are also in their infancy (Bastide at al. 2018). Here, we used phylogenomic data to infer the history of reticulation in baobabs and then developed and deployed network-aware ancestral state reconstruction methods to evaluate the possible role of adaptive introgression in explaining apparent floral homoplasy.

The baobab genus *Adansonia* (Malvaceae) includes eight morphologically distinct species (Fig. 1) (Baum 1995b; Cron et al. 2016). The group shows an unusual geographic
ANCIENT INTROGRESSION IN THE BAOBABS

distribution, with one tetraploid species (*A. digitata*) that is widespread across continental Africa, one diploid species (*A. gregorii*) endemic to Northwestern Australia, and six diploid species that are restricted to Madagascar. Previous phylogenetic analyses identified three distinct lineages corresponding to geography; however, relationships among these remain unresolved (Baum et al. 1998; Pettigrew et al. 2015).

Within Madagascar, the six recognized species are assigned to two groups based on morphology (Baum 1995b). Section *Brevitubae* includes a pair of allopatric species, *A. grandidieri* in the southwest and *A. suarezensis* in the north, which form a clade well-supported by ITS sequences (Baum et al. 1998). Both species are dry-season flowering and have flowers that are characterized by short staminal tubes, white petals, cream-colored inner calyces, and are inferred to be pollinated by fruit bats and lemurs (Baum 1995a; Andriafidison et al. 2006). The four Malagasy species of section *Longitubae* are wet-season flowering and share elongated flowers, with long staminal columns, red or yellow petals, and a reddish inner calyx and style. This floral morphology is associated with pollination by long-tongued hawkmoths (Baum 1995a; Ryckewaert et al. 2011). Despite their floral similarities, prior molecular phylogenetic analysis (Baum et al. 1998) has found limited support for the monophyly of Malagasy section *Longitubae*, and neither do they support a sister-group relationship between the Malagasy section *Longitubae* and the Australian species, which has elongated, white, primarily hawkmoth-pollinated flowers, and was formerly included in section *Longitubae* (Hochreutiner 1908; Perrier de la Bâthie, 1955; Baum 1995b). Prior, phylogenetic analyses do, however, support a clade composed of *A. madagascariensis*, *A. za*, and *A. perrieri* (hereafter referred to as “core Longitubae”), to the exclusion of *A. rubrostipa* (Baum et al. 1998). Despite differences in flowering phenology among sympatric baobab species in Madagascar, prior work has suggested
gene tree discordance in the core Longitubae (Baum et al. 1998) and the possibility of
introgression among species of Malagasy Section Longitubae (Leong Pock Tsy et al. 2013).

In order to infer explicit phylogenetic networks, meaning genealogical histories that
incorporate both incomplete lineage sorting (ILS) and reticulation (hybridization and
introgression), information from many independent gene genealogies are needed (Raymond et al.
capture, or hyb-seq, has increased in popularity as a source of hundreds of low-copy nuclear
genes for the purpose of multigene phylogenetics in non-model systems (Ekblom and Galindo
2011; Zimmer and Wen 2015; Grover et al. 2015, Harvey et al. 2016, Chau et al. 2018; Wolf et
al. 2018). Biotinylated RNA “baits” selectively enrich targeted genomic loci in next-generation
sequencing libraries. Furthermore, organellar and ribosomal sequences can usually be recovered
from the off-target reads (Weitemier et al. 2014). Targeted sequence capture is typically used
with short-read sequencing, which makes it appropriate and useful for samples with degraded or
poor-quality DNA, such as tissues from herbarium specimens (Hart et al. 2016, Villaverde et al.
2018).

Even when care is taken in bait design to target single-copy genes, the prevalence of
tandem, segmental, and whole genome duplications in plants (Van de Peer et al. 2009; Jiao et al.
2011; Wendel 2015; Conover et al. 2018) often leads to the joint recovery of paralogs.
Accordingly, caution needs to be exercised (Nicholls et al. 2015) in ortholog assignment and in
screening for chimeric sequences (Philippe et al. 2011; Struck 2013). Several bioinformatic
pipelines have been developed to facilitate assembly from hyb-seq data (Yang and Smith 2014;
ANCIENT INTROGRESSION IN THE BAOBABS

Here we used custom-designed baits to obtain sequences of hundreds of independent low-copy nuclear loci for all species of baobab. Despite the large size of our data set, relationships among the three geographic lineages could not be resolved. Within Madagascar the optimal trees support a sister group relationship between section Brevitubae and core Longitubae to the exclusion of A. rubrostipa. This implies floral homoplasy since core Longitubae and A. rubrostipa share elongated, yellow and red, hawkmoth-pollinated flowers, whereas section Brevitubae and A. digitata have short, white, mammal-pollinated flowers (A. gregorii has a medium length, white, mainly hawkmoth-pollinated flowers; Baum 1995a). Our data show that relationships in Adansonia are confounded by reticulation, with network inference methods supporting at least one major reticulation event, and possibly a second, though this inference is confounded by methodological limitations and read assembly challenges, possibly due to paralogy. We developed new phylogenetic comparative methods for explicit networks, in which discrete morphological traits may be inherited via reticulation. Using this method, we show that reticulation edges help explain floral evolution and apparent homoplasy in baobabs.

METHODS

Bait Design

Total RNA was extracted from fresh leaf tissue of Adansonia digitata L. and Bombax ceiba L., as in Chang et al. (1993), followed by cleanup using the Qiagen RNeasy kit following manufacturer instructions (QIAGEN Inc., Valencia, California, USA). RNA quality assessment was performed by Agilent RNA PicoChip Analysis (Agilent Technologies, Inc., Santa Clara,
California, USA) using 1µl of each sample diluted to 5 ng/µl. RNA library preparation was performed at the University of Wisconsin - Madison Biotechnology Center (Madison, WI) using 1 µg of total RNA with the Illumina TruSeq RNA Sample Prep kit (Illumina Inc., San Diego, CA, USA), followed by purification with Agencourt AMPure XP beads (Beckman Coulter, USA). Library quantification was checked with a Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific, USA) per manufacturer instructions. Samples were adjusted to a final concentration of 28-31ng/µl. Quality and quantity of the finished libraries were assessed using an Agilent DNA1000 chip (Agilent Technologies, USA) and Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific, USA), respectively. Libraries were standardized to 2nM. Cluster generation was performed using Illumina TruSeq Cluster Kits and the Illumina Cluster Station (Illumina Inc., San Diego, CA, USA). Paired-end, 100bp sequencing was performed at the University of Wisconsin - Madison Biotechnology Center (Madison, WI) using SBS chemistry on an Illumina HiSeq2000 sequencer. Images were analyzed using the Illumina Pipeline, version 1.8.2, and raw reads were assembled de novo with Trinity version 2.1.0 (Luo et al. 2012). The resulting contigs were used as BLAST queries against each other, the Arabidopsis ultra-conserved sequence database (http://cgpdb.ucdavis.edu/cgpdb2/), and the Gossypium exome (Paterson et al. 2012). For each contig showing reciprocal best-BLAST matches between Adansonia and Bombax, Adansonia-Bombax-Gossypium alignments were identified that had >800bp of continuously aligned sequences and an average pairwise sequence similarity ≥ 93.5% between the Bombacoids and cotton. These candidates were screened to remove candidate targets with repetitive sequences by RepeatMasker (Smit et al. 2015) and for potential gene families (sequences that clustered with two or more genes from the cotton transcriptome). When base calls were ambiguous in Adansonia, the bait sequence was based on Bombax. When both species
ANCIENT INTROGRESSION IN THE BAOBABS

were polymorphic at a position, the most common base call was selected. The resulting targets were used as the basis for the design and synthesis of 120 bp, 2X-tiled MYbaits (Arbor Biosciences, formerly Mycroarray, Ann Arbor, MI, USA), available as Supplementary File S1 (see Electronic Supplemental Material; https://doi.org/10.5061/dryad.2z34tmpgt). Raw reads were deposited in the Sequence Read Archive (SRA) of NCBI, BioProject SUB6460831. Transcriptomes are available on NCBI Sequence Read Archive under accession PRJNA493960 (Conover et al. 2018).

Taxon Sampling and Targeted Sequence Capture

Sampling included one to three accessions per *Adansonia* species in addition to three outgroups (Table S1). DNA was extracted from silica-dried leaf tissue or seeds with the Qiagen DNeasy Plant Mini Plant Kit (Qiagen, USA), following manufacturer instructions, but with the following modifications: (1) increased lysis buffer to 650µl and included 10µl Proteinase K (25mg/mL), (2) tissue and lysis buffer was incubated at 65°C for 20 minutes rather than the recommended 10 minutes, (3) all centrifugation steps were performed at 4°C, and (4) final elution used heated buffer (approx. 80°C) and was then incubated at room temperature for 10 minutes before centrifugation. DNA quality and quantity were estimated by 1% agarose gel electrophoresis. Qubit Fluorometric Quantitation (Life Technologies) was used for further quantification prior to DNA library preparation.

Illumina TruSeq DNA libraries were constructed and barcoded by the University of Wisconsin - Madison Biotechnology Center or the Genomics Core Facility at West Virginia University. Genomic DNA was sheared using a Covaris shredder to achieve a standard fragment
range of 500-600 bp. Sequence capture was performed at Iowa State University using the MYbaits protocol version 2 (Arbor Biosciences, formerly Mycroarray, Ann Arbor, MI, USA). Briefly, libraries were denatured and hybridized with biotinylated RNA capture baits over 36 hours. Quantity and quality of the captured libraries were assessed via Quan-it PicoGreen dsDNA assay (ThermoFisher Scientific, USA) and Agilent Bioanalyzer 2100, respectively. Enrichment of post-capture capture libraries was verified via QPCR as described previously (Salmon 2012, Grover 2017).

Target-enriched, Illumina TruSeq libraries for an initial twelve accessions (see Table S1) were pooled (index and sequence details are available at [Github](https://github.com/nkarimi/Adansonia_HybSeq)) and sequenced at the UW-Madison Biotechnology Center (Madison, WI) on a single lane of Illumina MiSeq with 300bp paired-end reads. An additional six accessions (see Table S1; Ama006, Ama018, Ape009, Aru127, Aza037, Aza135) were sequenced at 2x250bp by the Beijing Genomics Institute (BGI, Hong Kong) on an Illumina HiSeq2500 sequencer. Raw reads were quality trimmed using Trimmomatic v0.36 (Bolger et al. 2014) with the parameters $\text{ILLUMINACLIP}:\text{Adapters.fa}:2:30:15 \text{ LEADING}:28 \text{ TRAILING}:28 \text{ SLIDINGWINDOW}:8:28 \text{ SLIDINGWINDOW}:1:10 \text{ MINLEN}:65 \text{ TOPHRED33}$ (all scripts available at [Github](https://github.com/nkarimi/Adansonia_HybSeq))

**Read Assembly and Dataset Selection**

The first phase of assembly (see Supplementary Fig. S1 for flow chart) used the HybPiper package (Johnson et al. 2016) with our initial target set as a reference. HybPiper, yields a “paralog warning” whenever more than one contig is assembled that each covers >85%
ANCIENT INTROGRESSION IN THE BAOBABS

of the target. For putative single-copy targets, which yielded no paralog warnings for any of the
initial 12 (MiSeq-derived) accessions, we retained the resulting sequences and aligned them
using MAFFT version 7.299 (Katoh et al. 2012). Whenever HybPiper gave a paralog warning
for any of these taxa we retained all sequence variants for all accessions, aligned them using
MAFFT, and generated maximum likelihood trees in Geneious version 8.0.5
(http://www.geneious.com, Kearse et al. 2012) using the RAxML plugin. The resulting gene
trees were inspected to determine if the multiple contigs most likely represented alleles
(accessions from the same species form a clade) or paralogs (distinct clades each with multiple
species). For gene trees without evidence of paralogs, we retained a single allelic sequence for
each accession (the longest one, as selected by HybPiper).

Rather than discarding targets with paralogs, we sought to generate paralog-specific
references using gene tree-guided orthology identification. As summarized in Supplementary
Fig. S1, a separate consensus sequence was generated from all sequences of each putative
paralog and these paralog-specific consensus sequences were used as new, operational targets for
assembly in HybPiper. If the resulting assemblies lacked paralog warnings, and also appeared as
distinct clades on gene trees generated after alignment of all paralogs for a given original target,
each paralog assembly was retained. If sequences from the operational targets did not separate on
the resulting gene trees, suggesting orthology-paralogy mixing, the target was dropped. If,
alternatively, gene trees suggested yet further paralogs, the process was repeated iteratively until
each target was either dropped or yielded one or multiple alignments of putatively orthologous
sequences.

Preliminary network analyses suggested that Malagasy Longitubae samples sequenced on
the HiSeq platform might share erroneous signals of gene flow, perhaps due to the assembly of
Karimi, Grover, Ané, Gallagher, Wendel, Baum

distant paralogs that were not found using the shallower read depth of the MiSeq platform. Two
morphologically and geographically distinct species, *A. rubrostipa* and *A. perrieri*, were both
represented by one accession generated on the MiSeq platform (Aru001; Ape001) and one on
HiSeq (Aru127; Ape009). Therefore, we used PAUP* 4.0a (Swofford 2003) to identify all
targets whose optimal gene trees, as estimated by maximum likelihood trees in RAxML version
8.2.10 (Stamatakis 2006), failed to satisfy the unrooted backbone constraint ((Aru001,
Aru127),(Ape001, Ape009)). Removal of these genes resulted in our primary HybPiper data set
of 372 putative orthologs.

Although prior work has shown that allele phasing may have minimal impact on
phylogenetic inference (Kates et al. 2017), we sought to obtain allelic information as an
additional set of alignments. To generate clean targets for haplotype assembly, we used
consensuses of all targets (for all accessions) in the primary HybPiper data set after dropping a
further 28 targets whose optimal RAxML trees suggested the possibility of mis-assembly. This
included trees with single terminal branches at least three times longer than others on that tree, or
topologies that could not be rooted to support a monophyletic *Adansonia* clade. The 344
“cleaned” targets meeting these criteria were used to infer phased haplotypes. Trimmed reads
were mapped to the 344 accession-specific references using BWA v.0.7.15 (Li and Durban 2009)
with the *bwa mem* algorithm. Haplotypes were then inferred using HapHunt in BamBam (Page et
al. 2014) under the following parameters: (1) 5 runs per accession; (2) a minimum of 20x
coverage of each single nucleotide polymorphism (SNP); and (3) four haplotypes allowed for the
tetraploid *A. digitata* and two for all other taxa. We then generated 10 alternative HapHunt
alignments (HH1-10), each containing one haplotype sequence per target per accession, sampled
at random from the haplotypes inferred for that accession.
Species-Tree Phylogenetic Analyses and Network Inference

Bayesian phylogenetic inference was performed on each gene separately for all datasets with MrBayes (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003), either as a free-standing application or as implemented in the TICR pipeline (Stenz et al. 2015). MrBayes analyses used 4 linked chains with a heat of 0.2 and were run for 2,000,000 generations, with 25% discarded as burn-in. Convergence was assessed by the standard deviation of split frequencies. The resulting posterior distributions were analyzed in BUCKy version 1.4.4 (Ané et al. 2006, Larget et al. 2010) using alpha=1 and 1,000,000 generations. The nuclear HapHunt datasets, HH1-10, were analyzed individually. Additionally, to shed light on the effect of allele sampling, we combined the 10 posterior tree distributions for each gene into a composite posterior and used the resulting combined HapHunt data set for several downstream analyses.

Bayesian Concordance Analysis (Ané et al. 2007; Baum 2007) was implemented with BUCKy (Ané et al. 2007; Larget et al. 2011). BUCKy allows one to estimate the proportion of gene trees supporting a certain clade (the concordance factor, CF), while taking into account uncertainty in individual gene trees. BUCKy also estimates a population tree under the assumption that all discordance is due to ILS [we will use the term “population tree” rather than the more widespread term “species tree” so as to avoid the implication that the tips deserve recognition at the species rank]. BUCKy does not yield a support metric for the population tree per se, but does estimate a concordance factor (CF) for each clade, the percentage of gene trees having that clade, with credibility intervals. This allows us to identify clades whose CF is significantly higher than any conflicting clade and indicate those CFs with parentheses.
We also generated maximum likelihood gene trees with RAxML version 8.2.10 (Stamatakis 2006, 2014) and then used these to infer a population tree with posterior probabilities (PP) in ASTRAL-III (Mirarab et al. 2014; Zhang et al. 2018). Nuclear gene alignments were also used to infer a population tree with bootstrap support values using SVD quartets (Chifman and Kubatko 2015) in PAUP* (version 4a; Swofford et al. 2003).

To infer an explicit phylogenetic network we used the maximum pseudolikelihood method implemented in SNaQ (Solís-Lemus and Ané 2016), which infers reticulate evolutionary histories while accounting for ILS. The concordance factor table generated by BUCKy in the TICR pipeline was used as input into SNaQ. To test for the effect of input tree, starting population trees with branch lengths in coalescent units were obtained from both BUCKy and from the TICR pipeline, the latter of which uses Quartet MaxCut (Snir and Rao 2012).

We ran 50 independent runs of SNaQ for each dataset to infer the optimal network without hybridization edges (h0). This was then used as input for network searches with one hybridization edge (h1). Network searches were increased sequentially, up to 3 hybridization edges, in each case starting from the previous (h1 or h2) optimal network. The preferred number of hybridizations was determined based on analysis of the slope of a plot of log-pseudolikelihood against hybridization number (Solís-Lemus and Ané 2016); evaluation of 3 hybridization edges was deemed sufficient given the lack of change in the slope heuristic. The network with the best log-pseudolikelihood score for the optimal number of hybridizations (in each case one hybridization) was then selected as the starting network for bootstrap analysis. We generated a total of 100 independent bootstrap replicates (sampling from the confidence intervals for each possible quartet CF) with 20 runs per replicate, where 10 runs/replicate started with the optimal hybridization network and the other 10 started with the h0 population tree.
ANCIENT INTROGRESSION IN THE BAOBABS

Four-Taxon D-statistic (ABBA-BABA)

Guided by observed sensitivity of SNaQ to taxon sampling (see Results), we also used the D-statistic “ABBA-BABA” test (Kulathinal et al. 2009) on the concatenated gene alignments from the HybPiper dataset using a modification of the “CalcD” function from the R package evobiR (Blackmon et al. 2015) with ambiguous sites dropped. To test for a hybridization between *A. digitata* and section *Brevitubae* we analyzed all 162 BLdo quartets, namely those that contain one member each of *Brevitubae* (*A. grandidieri; A. suarezensis;*), core *Longitubae* (*A. madagascariensis; A. perrieri; A. za*), *A. digitata*, and the outgroup. Likewise, to test for a hybridization between *A. rubrostipa* and the core *Longitubae* we tested all 108 BLro quartets, namely those that contain one member each of *Brevitubae*, core *Longitubae*, *A. rubrostipa* and the outgroup. To test for sensitivity to taxon selection, we also used *A. gregorii* in place of the outgroups, yielding datasets BLdg and BLrg. For each quartet we calculated both the total number of ABBA and BABA sites in the concatenated alignment and the number of genes having more ABBA sites than BABA sites, or vice versa. Positive D-statistics show an excess of ABBA sites, while negative values show an excess of BABA sites. To evaluate significance, we used non-parametric bootstrap resampling of genes to obtain (for each 4-taxon data set) a Z-score, which is the calculated D-statistic divided by the bootstrap-estimated standard deviation of the D-statistic.

Plastome Assembly and Analysis
Reference guided assemblies of off-target plastid reads obtained after hyb-seq were performed using Burrows Wheeler Aligner (Li and Durban 2009) with the *bwa mem* algorithm and the *Gossypium raimondii* plastome sequence as reference (NCBI GenBank ID: HQ325744). Consensus sequences of the mapped reads were extracted from the BAM alignments and aligned with MAFFT (Katoh et al. 2012) using the FFT-NS-2 algorithm. We tested for recombination blocks within the plastomes using MDL (Ané 2011) with a minimum block length of one hundred parsimony informative sites. On each of the resulting partitions and the total concatenated alignment, jModelTest 2 (Darriba et al. 2012) was used to select a model of evolution, after which Bayesian phylogenetic inference was performed with MrBayes v3.2.3 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003). Analyses were run for 2,000,000 generations with 4 runs, 4 chains and a heat of 0.2 with 25% of generations discarded as burn-in. We also used RAxML version 8.2.10 (Stamatakis 2006, 2014) to infer maximum likelihood trees using the GTR-\(\Gamma\) model with 100 bootstrap replicates.

To test whether ILS, given the optimal network, could plausibly explain the recovered plastid phylogeny, we simulated expected plastid trees given two alternative population networks as inferred from the nuclear data. For each, we simulated 100,000 gene trees in the program hybrid-lambda (Zhu et al. 2015), after first multiplying all branch lengths in the network by 4 to account for the fact that plastid DNA experiences an effective population size one quarter that of nuclear genes. External branches were set to an arbitrary length of 1. Simulated tree topologies were input into PAUP* (version 4a; Swofford et al. 2003). We used tree filters to determine the frequency of targeted topologies among the 100,000 simulated gene trees.

*Phylogenetic Comparative Method for Discrete Traits on a Species Network*
To analyze flower color and pollinator mode in baobabs, we implemented maximum likelihood estimation of evolutionary parameters for discrete traits. We consider a trait with \( k \) possible states as evolving under a Markov process along each edge of a known species network. At a reticulation node in the network, we assume that the trait of the hybrid population, \( Y_h \), was inherited from either the major parent population, \( Y_1 \), or the minor parent population, \( Y_2 \), with probabilities equal to the proportion of genes contributed by each parent population (\( 1- \gamma > 0.5 \) and \( \gamma < 0.5 \)):

\[
Y_h = \begin{cases} 
Y_1 \text{ with probability } 1 - \gamma \\
Y_2 \text{ with probability } \gamma
\end{cases}
\]

This model may be interpreted in various ways. For instance, the trait might be controlled by a single gene, but this gene is unknown and probabilities \( 1 - \gamma \) and \( \gamma \) serve as prior probabilities that this unknown gene came from one parent or the other. Alternatively, the trait might be controlled by many genes of small effects, of which proportions \( 1 - \gamma \) and \( \gamma \) are expected to come from each parental population a priori. The \( \gamma \) inheritance probabilities (one \( \gamma \) for each reticulation) are assumed to be known, along with the network topology and branch lengths. At the root of the phylogenetic network, the prior probability of each state may be assumed to be uniform across the \( k \) states, or may be taken as the stationary distribution of the process given the transition rates. This model was already considered by Strimmer et al. (2001), who applied it to nucleotide data for the inference of ancestral recombination graphs. Like Strimmer et al. (2001), we calculate the likelihood of the trait data as a linear combination of likelihoods from each tree displayed in the network, while simultaneously optimizing transition rates. This implementation allows for non-symmetric rates; we used a likelihood ratio test to evaluate this asymmetric model against the nested model that enforces equal rates of trait gain and loss.
Conditional on the estimated rates, ancestral state estimations are obtained as the posterior probabilities of each state given the trait data at the leaves of the phylogenetic network. At each reticulation node, the posterior probability $p_{gf}$ that the trait was inherited via gene flow is calculated as the posterior probability that the trait state was inherited from the minor parent population (with inheritance $\gamma<0.5$) given the trait data at the leaves of the network and given the estimated rates. This posterior probability $p_{gf}$ is compared to the prior probability $\gamma$, to get a Bayes factor comparing the hypotheses of inheritance via gene flow from the minor parent versus vertical inheritance:

$$\frac{p_{gf} (1 - \gamma)}{\gamma (1 - p_{gf})}$$

Our implementation is available in the open source Julia package PhyloNetworks (Solís-Lemus et al. 2017), version 0.10.0.

**Ancestral Trait Reconstruction and Network Calibration**

To infer the most likely flower color and pollinator of ancestral populations, we considered a Markov process for the evolution of these traits on inferred phylogenetic networks, as described above. For pollinators, we considered a binary trait with states, hawkmoth-pollinated and mammal-pollinated, and for flower color we considered a binary trait with states, white and pigmented (=yellow or red). We scored the closest outgroup, *Scleronema micranthaum* as white flowered. Given that the methods cannot handle multiple states, *Scleronema* was dropped from the pollination analysis due to being interpreted as having mixed bat and moth pollination (van Dulmen 1998). Traits evolved along each lineage according to a 2-state Markov process with 2 transition rates and we assumed an equal prior probability for all states at the root of the phylogenetic network.
ANCIENT INTROGRESSION IN THE BAOBABS

The trait evolution model uses branch lengths in the network as a measure of evolutionary time. However, SNaQ infers branch lengths in coalescent units for internal branches, and does not infer any length for external branches (present-day populations). Therefore, we calibrated the branch lengths of the network using the approach developed by Bastide et al. (2018) and implemented in PhyloNetworks (Solís-Lemus et al. 2017). Briefly, we calculated the pairwise genetic distances between taxa from the individual gene trees, in which branch lengths measure the number of substitutions per site. To account for rate variation across loci, the tree for each locus was re-scaled to a common median patristic distance between outgroup taxa and ingroup taxa (all loci have at least one outgroup taxon and one ingroup taxon). Note that the total tree length was not used to normalize gene trees, because genes with missing taxa are expected to have a lower tree length due to missing branches. The observed pairwise distance matrix was calculated by averaging the pairwise distances across all loci, for each pair of taxa. Ages of nodes in the network were optimized to yield phylogenetic distances that matched the observed pairwise distances as well as possible, according to the ordinary least-squares criterion. Networks inferred from the primary HybPiper data set of 372 genes were calibrated using the pairwise distances obtained from the RAxML trees of these 372 genes. Networks inferred from the 344-gene HapHunt data set were calibrated using the pairwise distances obtained from the RAxML trees of one of the haplotype sets (HH6), whose network estimated with SNaQ has an estimated $\gamma$ closest to that of the combined HH1-10 network. Ancestral trait reconstructions were then performed independently using each calibrated network.

RESULTS
Targets and Dataset Selection

Analysis of transcriptomes recovered a total of 380 candidate sequences that met our criteria, for an initial total target space of 734,503 base pairs. The mean targeted locus was 1932.9 base pairs in length (minimum length of 763 bp and maximum of 7042 bp) with an average GC content of 42.8%. These targets are available as Supplementary File S1. An average of 9.3M and 17.9M read pairs were recovered for each accession sequenced on the MiSeq and HiSeq, respectively. MiSeq was run at 2x300nt and, thus, recovered an additional 100nt per read pair than HiSeq, which was run at 2x250nt. Nonetheless, even multiplexing 24 samples per HiSeq2500 (RapidRun) lane as compared to 12 samples per MiSeq lane, HiSeq still recovered an average of 3.4 Gb more sequence per sample. If all reads were on-target the average MiSeq and HiSeq coverage per nucleotide in the target would 5,000-fold to 12,000-fold, respectively. Even allowing for the fact that about half of the sequences are off-target, we expected all targeted genes to exceed 1000x coverage.

Following initial read mapping to the 380 gene bait set in HybPiper, we removed 45 targets and split some of the other targets into more than one discrete paralog. The resulting dataset included 412 genes, some of which seemed to cluster sequences by platform (MiSeq vs. HiSeq) rather than species. We infer that the increased coverage of sequencing in using HiSeq resulted in additional paralogous sequences being assembled for some HiSeq-based accessions, resulting in erroneous clustering of these accessions. Using the topological constraints described in the methods, we identified forty loci whose gene trees might have been distorted by a HiSeq-MiSeq artifact, which were subsequently removed, generating our primary HybPiper dataset of
ANCIENT INTROGRESSION IN THE BAOBABS

372 genes. These 372 genes were from 241 original targets which yielded single copy assemblies and 55 original targets yielded multiple paralogs. Of the latter, 53 were split into two paralogs (in five cases we did not retain both copies), five original targets were split into three paralogs, and one target was split into each of four, five, and six paralogs. In total 131 paralogous alignments were included in the 372-gene HybPiper data set.

After dropping a further 28 targets, as described in Methods, 344 modified targets were used to call haplotypes with HapHunt. Consistent with baobabs being outcrossing (Baum, 1995a; Venter et al. 2017), haplotype recovery for the diploid *Adansonia* taxa resulted in an average of 73.5% loci inferred to be heterozygous (i.e., with two variant sequences per accession). Across the three accessions of tetraploid *A. digitata*, 1.5% of loci were homozygous for a single allele, 50% yielded two variants, 12.5% yielded three, and 36% recovered four (Fig. S2).

Population Tree Inference

Results from all three population tree estimation methods and both data sets were in agreement with the exception of the SVDQuartets tree from the HybPiper dataset (Fig. S3a), which placed *A. rubrostipa* sister to the *Brevitubae* taxa (BS=59) and placed *A. perrieri* as sister to the rest of the core Longitubae (BS=93). Using the other methods, the optimal population trees inferred for the HybPiper and HapHunt datasets were identical except in their resolution of the basal node within *Adansonia* (Fig. 2; Fig. S3). Whereas the HybPiper data placed *A. gregorii* sister to the rest of the genus (PP=0.50; CF=0.19 n.s.), the HapHunt data placed it sister to *A. digitata* (PP=1, CF=0.37). In other regards, the optimal population tree is robust to assembly method. Furthermore, the topology matches the concordance tree constructed by BUCKy. The
concordance factors are generally low, suggesting high levels ILS, but generally higher for the HapHunt data. The most dramatic difference relates to the monophyly of the Malagasy clade whose CF is 0.50 for the HapHunt data, as contrasted with CF=0.27 for the HybPiper data set.

On the optimal tree for both datasets, the Malagasy members of section Longitubae do not form a clade. Instead, *A. rubrostipa* is sister to a clade including section Brevitubae plus core Longitubae (*A. madagascariensis, A. perrieri, and A. za*). Although the section Brevitubae-core Longitubae clade has a concordance factor of only 0.19 based on the HybPiper data, the corresponding credibility interval (0.169 - 0.210) does not overlap the CF of a four-species Longitubae clade (0.116 - 0.151). In contrast, the section Brevitubae-core Longitubae clade in the summed HapHunt dataset has a CF=0.23 whose credibility interval (0.198 - 0.253) does overlap that of 4-taxon Longitubae clade (0.151 - 0.206). The section Brevitubae plus core Longitubae clade is supported in ASTRAL analyses of the HybPiper and HapHunt datasets (posterior probability of 0.64 and 0.66, respectively).

The consistent signal contradicting the monophyly of Malagasy section Longitubae is surprising, given that the four species have very similar, elongated red-yellow flowers, which are quite different from the shorter, white flowers of Brevitubae and the outgroups. As a result, these data suggest either homoplasy of floral traits or that genes have discordant histories, whether due to ILS, introgression, or both.

The three species of core Longitubae (*A. madagascariensis, A. perrieri, and A. za*) are supported as a clade, but the two accessions of *A. za* are consistently resolved as non-monophyletic, with one *A. za* accession (from southern Madagascar) being sister to all other sampled core Longitubae (all of which are from northern populations), including the other *A. za*
ANCIENT INTROGRESSION IN THE BAOBABS

accession. Similar non-monophyly of *A. za* was reported based on ITS analysis (Baum et al. 1998).

*Phylogenetic Network Inference*

Based on the slope heuristic, SNaQ analyses favored h1 networks with a single reticulation edge for all datasets, with $\gamma$, the proportion of genes inferred to have followed the hybridization edge, ranging from 7-23%. It is not surprising that the h1 network is supported since, when we allowed 2 or more reticulation edges, higher order reticulation events had estimates of $\gamma < 2\%$. It should be noted, however, that SNaQ is constrained to ignore intersecting hybridization scenarios, meaning that once a preferred introgression edge is added, many other potential edges are not able to be recovered in the search (Solís-Lemus and Ané 2017).

With all taxa included, the HybPiper data supported introgression from the African lineage (the stem lineage of *A. digitata*) into the stem lineage of *Brevitubae*, with $\gamma = 12.4\%$ (Fig. 3a). This hybridization edge was recovered in 73% of the bootstrap replicates. Reducing this data set to those genes that were used for haplotype inference, yielded the same network (as in Fig. 3a), with the hybridization edge being found in 95% of bootstrap replicates and $\gamma = 8\%$.

The HapHunt dataset with all taxa included also identified (in 70% of bootstrap replicates) an introgression between *Brevitubae* and a non-Malagasy lineage (Fig. 3b). However, the direction of introgression is reversed, suggesting introgression of 7.2% of genes from *Brevitubae* into a common ancestor of *A. digitata* and *A. gregorii*. Interpreted literally, this scenario is extremely unlikely for geographic and temporal reasons. It would suggest, for
example, gene flow from an ancestor of section *Brevitubae* on Madagascar into an African ancestral population prior to long-distance dispersal of *A. gregorii* to Australia, making the common ancestor of *A. gregorii* and *A. digitata* significantly younger than the radiation of the Malagasy clade, which is at odds with prior molecular dating analyses (Baum et al. 1998).

Given the difficulty of resolving deep splits and the fact that directionality sometimes cannot be determined reliably with SNaQ, (Solís-Lemus and Ané 2017), we explored the sensitivity of these results to taxon sampling. In the process we discovered that deletion of the outgroups suggested a different reticulation history. For both the HybPiper and HapHunt datasets, the optimal networks after pruning outgroups implies introgression (γ = 13-17%) between *A. rubrostipa* and the stem lineage of the core Malagasy Longitubae clade (Fig. 3c,d).

Given the impossibility of inferring intersecting hybridization cycles with SNaQ, we sought to evaluate the possibility of additional introgression events after deleting selected taxa (Fig. S5). Taxon selection was guided by the plastid tree, and designed to include representatives from each of the primary clades to help integrate the resulting networks. Including an outgroup but deleting Brevitubae taxa from the HybPiper data set supported gene flow from *A. rubrostipa* to the southern *A. za* sample, with γ = 19.5% (Fig. S5d). In contrast, deleting *A. rubrostipa* supported reticulation between *A. digitata* or *A. digitata*+*A. gregorii* and Brevitubae (γ = 11-15%; Fig. S5a,c), as well as additional gene flow between northern and southern *A. za* accessions (γ = 42%, Fig. S5b). Taken together the nuclear data provides evidence of introgression between *A. rubrostipa* and core Longitubae or from an African lineage into section *Brevitubae*, plus the possibility of additional gene flow between accessions of *A. za*.

Following analysis of each of the ten individual HapHunt datasets, six yielded results consistent with the combined HapHunt dataset (Fig. S4) only differing in the hybridization
fraction, $\gamma$, which ranged from 7.2 - 8.5%. Of the remaining four, two indicated a hybrid edge between *A. digitata* and one member of Brevitubae, *A. suarezensis*, similar to the combined HapHunt dataset when *A. gregorii* was pruned (not shown). One out of the ten HapHunt datasets yielded a network similar to the one obtained without outgroups, which involves gene flow from *A. rubrostipa* to the core Longitubae clade ($\gamma=13.2%$; Fig. S4, I).

**D-statistic (ABBA-BABA)**

Depending on taxon-sampling, SNaQ yielded one of two reticulation edges: between *A. digitata* and Brevitubae, or between *A. rubrostipa* and core Longitubae. On an unrooted network (Fig. 4), these two scenarios are quite close and form intersecting loops. This means that SNaQ can never yield support for both occurring on the same network, leaving us unsure if one edge is true, with the other being an artifact of some kind, or if both occurred.

To evaluate these two hypotheses separately we conducted ABBA-BABA tests targeted at either of the two hybridizations individually. To test the *A. digitata*-Brevitubae edge, we dropped *A. rubrostipa* and looked for polymorphisms supporting each partition of the four remaining groups, B, L, d, and o/g (Fig. 5). Likewise, we analyzed the BLro and BLrg datasets to test for gene flow between *A. rubrostipa* and core Longitubae. Support for the reticulation events would come from an excess of BABA over ABBA sites in BLdo/BLdg and the reverse under BLro/BLrg.

As summarized in Figure 5a, summing over both the HybPiper and HapHunt datasets, only BLrg quartets yielded statistically consistent deviations from null expectations, supporting
positive values of $D$ and, consequently indicating introgression between core Longitubae and *A. rubrostipa*. The support for this conclusion is especially consistent for the HapHunt data set, which yields a positive Z-score, regardless of which set of 4 taxa are included in the test (Fig. 5b).

*Plastid-Nuclear Incongruence Suggests Additional Reticulation Events Within Madagascar*

Reference-guided assembly yielded nearly complete plastid genomes for each accession (number of reads assembled and read coverage is reported in Supplementary Table S2), with a shared alignment of 163,590 bp, containing 167-4,795 pairwise SNPs among taxa. Although it is generally assumed that the whole plastome has a single phylogenetic history, we first used MDL (Ané 2011) to identify possible recombination breakpoints. This recovered four partitions, representing 39.2, 21.7, 13.1, and 89.6 kilobases, respectively. All but the third partition (13.1 kb) supported a combined Australian and African clade (as seen in the HapHunt data set). Three partitions supported *A. suarezensis* as sister to the remaining Malagasy species, while partition two (21.7 kb) was unable or reject this arrangement based on bootstrap and posterior probabilities (Fig. S6). As the conflicts among the plastid partitions are relatively minor and plastid recombination is unlikely, we used the concatenated plastid tree to represent the plastid history (Fig. 6).

The plastid data strongly support a clade composed of the African and Australian species. The tree also implies non-monophyly of Brevitubae, with *A. suarezensis* strongly supported as sister to the rest of the Malagasy clade. It is worth noting that the plastid marker analyzed by Baum et al. (1998), *rpl16*, also supported a clade composed of all Malagasy species except *A.
ANCIENT INTROGRESSION IN THE BAOBABS

*A. suarezensis*. This is surprising given the many morphological similarities shared between *A. suarezensis* and *A. grandidieri*, including a distinctive crown architecture, white, upright flowers with short-staminal tubes, winter flowering, large seeds, and cryptocotylar germination (Baum 1995a). The plastid tree is also discordant with the nuclear population tree in supporting the monophyly of the four Malagasy *Longitubae* (*A. rubrostipa*, *A. za*, *A. perrieri*, and *A. madagascariensis*).

To evaluate whether nuclear-plastid discordance (Fig. 2 & Fig. 6) can be explained by ILS on the nuclear-derived network, we simulated 100,000 plastid trees on the optimal networks for the 372-gene HybPiper dataset, either with all taxa included (network with *A. digitata*-Brevitubae introgression) or with outgroups excluded (network with *A. rubrostipa*-core Longitubae introgression). Only one simulated tree (out of 100,000) supported *A. suarezensis* as sister to the rest of the Malagasy clade when using the *A. digitata*-Brevitubae introgression network, and four when using the *A. rubrostipa*-core Longitubae introgression network. Despite suggesting rarity of this topology, it is noteworthy that a clade of all Malagasy species except *A. suarezensis* was found in nine maximum likelihood gene trees, out of the 372 genes (2.4%) in the primary HybPiper data set, with three of these nine trees placing *A. suarezensis* sister to the rest of the Malagasy taxa.

The plastid tree also differs from the nuclear population tree in supporting southern *A. za* as sister to all Malagasy *Longitubae*, including *A. rubrostipa*. None of the simulated trees under the *A. digitata*-Brevitubae introgression have this resolution, but this topology is found in four trees simulated under the *A. rubrostipa*-core Longitubae introgression network. Similarly, among the 372 individual maximum likelihood gene trees inferred from the primary HybPiper dataset, four (~1%) showed all Malagasy *Longitubae* (including *A. rubrostipa*) as a clade sister to
southern A. za. In summary, while there is plastid-nuclear discordance under both networks, the plastid data are easier to reconcile with A. rubrostipa - core Longitubae introgression than with A. digitata-Brevitubae introgression.

Ancestral Trait Reconstruction: Introgression Explains the Shift in Pollination Syndrome in Brevitubae

Given that the estimated population tree shows non-monophyly of Malagasy Longitubae, which includes all the colored, hawkmoth pollinated species, we sought to explore the possibility that reticulation edges could help explain the evolution of floral morphology and pollination systems in Adansonia. We conducted ancestral state reconstruction of flower color and inferred pollination mode on time-calibrated networks. Time-calibrated networks from the HapHunt data featured many collapsed internal branches, reflecting temporally improbable introgression edges, especially that representing gene flow between the common ancestor of A. gregorii and A. digitata and Brevitubae.

For each model and trait, AIC favored equal rates of gain vs. loss, most likely due to the small sample size of the trait matrix. With these parameters, there was consistent support for introgressed genetic material having played a role in floral evolution (Fig. 7; Table S3). A Bayes factor (Bf) over 1 indicates positive evidence that a trait was acquired from the minor rather than from the major parent at a reticulation node. By this criterion, focusing on the HybPiper networks, there is strong evidence of flower-color introgression, whether it be of pigment from A. rubrostipa to core Longitubae on the network without outgroups (Bf = 7.4) or of non-pigment from A. digitata to Brevitubae on the network with outgroups (Bf = 8.0). There is also some
ANCIENT INTROGRESSION IN THE BAOBABS

support, with a Bayes factors of 2.4-3.0, for introgression introducing hawkmoth pollination into core Longitubae (no outgroups) or mammal pollination into Brevitubae (with outgroups). In either case, these data suggest that introgressive hybridization has transferred ecologically important traits between lineages. It is important to underscore that our model accounts for the likelihood of convergent evolution when integrating over all the unknown ancestral states (which will include convergent scenarios) to calculate the overall likelihood of the tip data. Nonetheless, comparing the fit of the trait data on each network versus its major tree (with the reticulation removed) yielded consistent results (Table S3).

DISCUSSION

Inference of Reticulation with Hyb-Seq data

Our data support a history of reticulation in Adansonia but, despite using a custom set of baits for 380 nuclear loci, the detailed reticulation history proved sensitive to assembly method and taxon inclusion. Recently, Lambert et al. (2019) also found that detection of introgression was sensitive to sequencing batch. Deeper sequencing, as achieved in our data set with the HiSeq platform used on some accessions, has been shown to improve detection of hidden paralogy (Philippe et al. 2011). Given this, we suspect that using a mixture of accessions sequenced with different platforms resulted in different paralogs being assembled for some accessions. To overcome these difficulties, we developed a thorough but laborious procedure of iterative assembly and tree examination to arrive at a conservative, manually curated data set.
Even after careful curation, conflicting reticulation histories were suggested depending on whether we did or did not include outgroups: analysis of all taxa supports a Brevitubae - *A. digitata* hybridization network, suggesting gene flow from Africa to Madagascar, whereas exclusion of the outgroups supports an *A. rubrostipa* – core Longitubae network, and hence introgression within Madagascar. Based on ABBA-BABA tests, concordance with the plastid data, and geographical proximity, we believe that exclusion of outgroups yields a more plausible result. This highlights the need to carefully examine alternative assemblies and taxon inclusion sets when using Hyb-Seq data for network inference.

**Biogeography of Adansonia**

In this study, despite obtaining hundreds of low-copy nuclear loci, relationships among the three primary geographic lineages of *Adansonia* could not be resolved. This begs the question as to the geographic origins of the genus. Previous molecular dating (Baum et al. 1998) rules out the possibility of a Gondwanan origin leading to extant distributions. Given that the sister lineage to *Adansonia* is Neotropical (Carvalho et al. 2016), it was previously suggested that the stem lineage of *Adansonia* migrated across the Atlantic to Africa (Baum, 1998), perhaps as part of the Boreotropical migration route that was open during periods of warmer climate in the Eocene (Tiffney 1985). This hypothesis is supported by the presence of Bombacoid pollen fossils in North America (Wolfe 1975) and Europe (Krutzsch 1989). However, the alternative of trans-Pacific dispersal into Australia and subsequently along the Indian Ocean rim needs also to be considered given the presence of bombacoid pollen fossils in Antarctica during the Eocene (Pross et al., 2012). Whereas the trans-Atlantic hypothesis is compatible with any resolution of
ANCIENT INTROGRESSION IN THE BAOBABS

the three geographic lineages of *Adansonia*, the trans-Pacific hypothesis predicts, by parsimony, that the African and Malagasy taxa would be sister to one another. The possibility of such long-distance dispersal events could be explained by baobab’s buoyant fruits that are potentially well-suited to hydrochory (Cornu et al. 2014) and have been shown to have viable seeds after extended seawater immersion (Leong Pock Tsy et al. 2009).

The nuclear data do not resolve the basal relationships of *Adansonia*: branch lengths are short and different assembly methods favor (albeit very weakly) different resolutions. The plastid data give strong support for an Africa+Australia clade, thus contradicting trans-Pacific dispersal. While one might attribute this to incomplete lineage sorting on the short internal branches of the inferred population trees, such a scenario would imply that multiple lineages carrying divergent plastid haplotypes dispersed across the Indian Ocean to Africa/Madagascar, which seems unlikely. Instead it seems more probable that there was an almost simultaneous divergence into the three extant lineages somewhere in northwest Africa or the Middle East, followed by eastward dispersal to Australia. Nonetheless, pending datasets with additional *A. gregorii* accessions and fewer paralogy-calling uncertainties, we currently consider the biogeographic history of baobabs to be unresolved.

*Introgression between the African lineage and Brevitubae by Overwater Dispersal*

While less certain than the *A. rubrostipa* – core Longitubae introgression, our analyses suggest the possibility of gene flow from an ancestor of *A. digitata*, presumably living in continental Africa, into an ancestor of Malagasy *Brevitubae*. While this conclusion could be an artifact driven by inclusion of relatively distant outgroups, and is not strongly supported by
ABBA-BABA tests, it is certainly plausible that there was more than one dispersal event of baobabs to Madagascar. Many lineages endemic to Madagascar are of African origin (Haber et al. 2017; Yoder et al. 1996) and multiple long-distance dispersal events to Madagascar within a single lineage are known (Kainulainen et al. 2017).

Networks that support *A. digitata* – *Brevitubae* introgression suggest that perhaps 10% of the genome of extant *Brevitubae* descended from *A. digitata*. This suggests that a single tree established in Madagascar, became reproductively mature, crossed with a local population and generated hybrid genotypes that were sufficiently favored (perhaps by mammal pollinators) that a significant fraction of the recipient species' genome was replaced. Alternatively, a population of *A. digitata* could have established in Madagascar before hybridizing with resident species and then was extirpated. While there are few clear cases of transoceanic dispersal and hybridization, this has famously been shown in cotton for which the allopolyploid cotton lineage of the Americas resulted from trans-oceanic dispersal of an A-genome taxon from Africa or Asia into the New World followed by hybridization with an indigenous American D-genome diploid (Wendel and Grover 2015, and references therein).

If introgression occurred from the African lineage (*A. digitata*) into the Malagasy clade, it must have been quite ancient. It presumably predated the origin of tetraploidy in *A. digitata* and the transition of the Malagasy *Brevitubae* to dry season flowering or its divergence into its two extant species (*A. grandidieri* and *A. suarezensis*). If the two lineages occurred in sympatry, gene exchange is plausible despite the main pollinators being different; based on extant species, nocturnal primates and hawkmoths would occasionally visit both flower types (Baum 1995a). To date, no artificial crossing studies have been done testing interfertility.
ANCIENT INTROGRESSION IN THE BAOBABS

Ancient Introgression between A. rubrostipa and Longitubae

Plastid and nuclear discordance suggests a history of gene flow between A. rubrostipa and core Longitubae. This conclusion is supported by SNaQ analyses that exclude outgroups and gains strong statistical support from ABBA-BABA tests. Given that A. rubrostipa and the other extant Longitubae share similar flower morphologies and pollination systems, and also have widely overlapping ranges and some potential for occasional flowering season overlap, such hybridization is plausible. Indeed, there are reasons to infer that there was not just one ancient hybridization event but several instances of gene flow between A. rubrostipa and core Longitubae both before and after the divergence of the latter into its three extant species. Such gene flow is in agreement with analyses based on nuclear microsatellite data (Leong Pock Tsy et al. 2013). Furthermore, a specific A. rubrostipa - A. za hybridization was identified in a h1 search after pruning Brevitubae taxa, though with lower likelihood than the corresponding h1 A. digitata - Brevitubae network. Due to intersecting cycles (shared branches) an A. rubrostipa - A. za reticulation edge cannot be found by SNaQ on an h2 network with either an A. digitata – Brevitubae or an A. rubrostipa – core Longitubae edge. Reticulation between A. rubrostipa and A. za was recovered, however, when used as the input network for the starting search. This also agrees with our unpublished data documenting additional reticulation events among the Malagasy Longitubae, all of which are wet-season flowering and primarily hawkmoth pollinated.

Plastid-Nuclear Tree and the Plastid Non-Monophyly of Brevitubae
Karimi, Grover, Ané, Gallagher, Wendel, Baum

The primary population tree derived from nuclear genes conflicts markedly with the plastid tree in the placement of two species, *A. suarezensis* and *A. rubrostipa*. Cases of plastid-nuclear tree discordance are commonly attributed to “chloroplast capture” (Rieseberg and Soltis 1991; Tsitrone et al. 2003; Feliner et al. 2017), which is to say introgression affecting the plastid but not (much of) the nuclear genome. Since we reconstructed an explicit network from the nuclear data, rather than just a single tree, we could assess whether the same hybridization history, combined with incomplete lineage sorting, could explain plastid-nuclear discordance.

The plastid data provides strong support (100% bootstrap and 1.0 PP) for the placement of *A. suarezensis* as sister to the rest of the Malagasy baobab clade. Simulations suggest that this resolution is unlikely, with a probability of $<4 \times 10^{-5}$, given neutral evolution and ILS along the nuclear-inferred networks. Nonetheless, the nuclear genes themselves suggest that such a signal also exists in the nuclear genome, with 2% of nuclear gene trees placing *A. suarezensis* as sister to the rest of the Malagasy clade. This hints at a possible undetected minor introgression event, which we might describe as an “introgressive kiss,” either involving the *A. grandidieri* lineage and the stem lineage of the Malagasy *Longitubae* or between *A. suarezensis* and a now extinct sister lineage to the entire Malagasy clade. While such additional hybridizations seem relatively unlikely given the long stem lineage of *Brevitubae* (i.e., *A. suarezensis* and *A. grandidieri*) and their synapomorphy of winter-flowering, it cannot be ruled out.

The plastid tree is also distinct in embedding *A. rubrostipa* within the core Malagasy *Longitubae* and placing southern accessions of *A. za* sister to the remainder of the *Longitubae* clade. This topology has a low probability of arising simply by ILS, especially on networks with a *Brevitubae-A. digitata* reticulation. Consequently, this result further supports the possibility of additional undetected reticulation events within the *Longitubae* clade.
ANCIENT INTROGRESSION IN THE BAOBABS

*Adaptive Introgression of Pollination Traits*

We developed methods for time-calibrating networks, inferring the probabilities of alternative character states at ancestral nodes, and then estimating the posterior probability that the trait was acquired from the minor or major parent during hybridization rather than as a result of homoplastic evolution. These methods should have broad applicability for studying character evolution in the context of phylogenetic networks. As an example of their application we used ancestral trait reconstruction analysis to explore whether adaptive introgression of floral pigment and pollination mode might explain the implied homoplasy associated with these traits when mapped onto the dominant population tree.

Considering, first, the best supported network, namely that entailing introgression from *A. rubrostipa* into a common ancestor of the three species of core Longitubae, we found strong support for the hypothesis that introgression transferred the shared floral traits of a red style, red interior calyx and colored petals from the *A. rubrostipa* lineage into Longitubae. Furthermore, as shown by analysis of pollination mode evolution on this network, a consequence of this introgression was likely to include a shift towards hawkmoth pollination. Indeed, it is plausible that adaptive introgression of a suite of floral traits, including not just pigmentation but staminal tube length, nectar composition, and scent, could have been driven by them promoting the frequency and efficiency of pollinator visits by hawkmoths.

Although less well-supported by our data, the *A. digitata – Brevitubae* gene flow hypothesis also supports a role for introgression in floral homoplasy. In this case, the highest likelihood reconstructions imply an introgression-mediated reversal in *Brevitubae* to an all-white
flower and a concomitant switch to mammal pollination. Although not formally analyzed here, we presume that introgression would also explain similarities between the donor and recipient taxa in shared traits such as nectar volume, nectar chemistry, and scent profile. On the other hand, to explain some striking differences in flower form between *Brevitubae* and *A. digitata*, including the latter’s long-pendulous flowers and extremely wide, reflexed petals, one could either suppose that these traits evolved in *A. digitata* after the introgression event or that these traits were present but failed to introgress.

While examples of introgression facilitating pollinator shifts have been reported (i.e. Louisiana irises, Wesselingh 2006; Monkeyflowers, Stankowski and Streisfeld 2015), this is the first case we are aware of that involves transitions between hawkmoth and mammal pollination. In the future it would be exciting to look more broadly at the genomes of all *Adansonia* species in the hopes of identifying candidate genes for various traits involved in the observed pollination syndromes, such as flower color, nectar characteristics or floral scent chemistry. Such genome-scale work would not only solidify the history of ancient introgression in the baobabs, but could provide a model for understanding mechanisms involved in introgressive pollination system evolution.

FUNDING

This project is based on work supported by the National Science Foundation award DEB-1354268.

ACKNOWLEDGMENTS
ANCIENT INTROGRESSION IN THE BAOBABS

We thank Steve Goldstein, Noah Stenz, and the Center for High Throughput Computing at the University of Wisconsin- Madison for computational assistance and Claudia Solís-Lemus who advised on network analyses. Computational support at Iowa State University was provided through the ResearchIT Unit (https://researchit.las.iastate.edu/). Ben Wirth, George Brown, and Diana Mayne, and the Darwin Botanic Gardens kindly provided samples. Alison Dawn Scott conducted RNA extractions, and the University of Wisconsin Biotechnology Center DNA Sequencing Facility provided RNA-Seq library preparation and sequencing services and assistance with bait design.

SUPPLEMENTARY MATERIAL

Supplementary material, including data files and/or online-only appendices, can be found in the Dryad data repository: https://doi.org/10.5061/dryad.2z34tmpgt, and custom scripts on GitHub: https://github.com/nkarimi/Adansonia_HybSeq. Raw reads were deposited in the Sequence Read Archive (SRA) of NCBI, BioProject SUB6460831.

REFERENCES


ANCIENT INTROGRESSION IN THE BAOBABS


Karimi, Grover, Ané, Gallagher, Wendel, Baum


ANCIENT INTROGRESSION IN THE BAOBABS


Karimi, Grover, Ané, Gallagher, Wendel, Baum


ANCIENT INTROGRESSION IN THE BAOBABS


Karimi, Grover, Ané, Gallagher, Wendel, Baum


Karimi, Grover, Ané, Gallagher, Wendel, Baum


ANCIENT INTROGRESSION IN THE BAOBABS


ANCIENT INTROGRESSION IN THE BAOBABS


ANCIENT INTROGRESSION IN THE BAOBABS


FIGURE CAPTIONS:

Figure 2. Population trees from (a) the primary HybPiper dataset with 372 genes and (b) the combined HapHunt dataset with 344 genes. Outgroups not shown. BUCKy and ASTRAL yielded the same tree topologies. BUCKy concordance factors added above branches (numbers in parentheses are not significant, having credibility intervals that overlap with any conflicting clades); ASTRAL posterior probabilities added below. Internal branches scaled with branch lengths in coalescent units (as estimated by BUCKy).

Figure 3. Phylogenetic networks as inferred by SNaQ from the HybPiper dataset (left) and HapHunt dataset (right), with all taxa included (a & b) or after deleting outgroups (c & d). Branch lengths scaled in coalescent units and minor inheritance probabilities, $\gamma$, shown in grey.

Figure 4. Scenarios for intersecting cycles given an unrooted phylogenetic network for all accessions (top) or all accessions minus outgroups (bottom).

Figure 5. Summary of ABBA-BABA tests. (a) Effect of quartet choice on the distribution of Z-scores. (b) Z-scores for BLrg quartets, color- and shape-coded by the taxon selected from section *Brevitubae*, with taxa from core Longitubae on the x-axis. See Supplemental Data for details.
ANCIENT INTROGRESSION IN THE BAOBABS

Figure 6: Concatenated plastid tree inferred by maximum likelihood and Bayesian phylogenetic inference. All branches have posterior probabilities (PP) of 1.0 and bootstrap support (BS) of 100%, unless otherwise indicated (PP/BS). Outgroups removed from figure.

Figure 7. Ancestral state reconstruction of pollination system (top) and flower color (bottom) on time-calibrated networks inferred from the HybPiper nuclear dataset. Analyses on the left were based on networks with outgroups included, although these outgroups were excluded from analysis of pollination mode because *Scleronema* has a mixed pollination system (see Methods).

SUPPLEMENTAL MATERIAL

Data S1. Targeted sequence capture baits file.

Figure S1. Read assembly pipeline.

Figure S2. Haplotype recovery heatmap.

Figure S3. SVDQuartets tree from the (a) HybPiper and (b) HapHunt datasets.

Table S1. Taxon sampling.

Table S2. Plastid genomes assembly statistics.

Table S3: Model summary.
Figure 1. Eight species of Adansonia. A) A. digitata, continental Africa, B) A. gregorii, Australia, C) A. grandidieri, Madagascar, D) A. suarezensis, Madagascar, E) A. madagascariensis, Madagascar, F) A. perrieri, Madagascar, G) A. za, Madagascar, H) A. rubrostipa, Madagascar.
Outgroups

A. gregorii 01

A. digitata 01
A. digitata 02
A. digitata 03

A. rubrostipa 01
A. rubrostipa 127

A. suarezensis 01
A. grandidieri 01
A. grandidieri 02

A. za 135
A. perrieri 01
A. perrieri 09
A. za 037

A. madagascariensis 06
A. madagascariensis 018

12.4%

7.2%

17%

12.8%

1
Figure 5. Summary of ABBA-BABA tests. (a) Effect of quartet choice on the distribution of Z-scores. (b) Z-scores for BLrg quartets, color- and shape-coded by the taxon selected from section *Brevitubae*, with taxa from core *Longitubae* on the x-axis. See Supplemental Data for details.