**Phylogenomic analysis of an unusual biogeographic disjunction in the cotton tribe (Gossypieae)**

Corrinne E Grover, Mark A Arick II, Justin L. Conover, Adam Thrash, Guanjing Hu, William S Sanders, Rubab Naqvi, Muhammad Farooq, Xiaochong Li, Lei Gong, Joann Mudge, Thiru Ramaraj, Joshua A Udall, Daniel G Peterson, Jodi Scheffler, Brian Scheffler, and Jonathan F Wendel

**Abstract**

**Introduction**

One of the intriguing evolutionary phenomena that characterizes the cotton tribe, *Gossypieae*, is the prevalence of long-distance, trans-oceanic dispersal {Wendel, 2015 #126}*.* The most famous of these are in the cotton genus (*Gossypium*), which includes the intercontinental dispersal of an African species to the Americas in the mid-Pleistocene {Wendel, 1989 #14} that gave rise to the New World allopolyploid cottons (including the primary cottons of commerce). Outside of *Gossypium*, multiple long-distance dispersals have occurred during the evolution of the tribe {Dejoode, 1992 #1;Fryxell, 1979 #11;Stephens, 1958 #12;Stephens, 1966 #13;Wendel, 1989 #14;Wendel, 1992 #15;Wendel, 1990 #16;Wendel, 1990 #17;Wendel, 2003 #18;Seelanan, 1997 #19}. One example includes the sister genera *Kokia* and *Gossypioides*, from Hawaii and southeast Africa, respectively. Based on molecular divergence estimates derived from both chloroplast and nuclear genes, these two genera are estimated to have diverged from each other in the Pliocene, approximately 3 million years ago (mya), and from *Gossypium* during the Miocene, perhaps 10-15 mya {Seelanan, 1997 #19;Cronn, 2002 #26}.

*Kokia* (Malvaceae) is a small genus of Hawaiian endemics comprising four species that were once widespread components of Hawaiian forests yet now are either endangered (three species) or recently extinct (*K. lanceolata* Lewton){Morden, 2017 #125}{Bates, 1990 #20;Sherwood, 2014 #21}). Few individuals remain of the two extant species, *K. kauaiensis* (Rock) Degener & Duvel and *K. drynarioides* (Seem.) Lewton, the latter being nearly extinct in the wild, while the third endangered species, *K. cookei* Degener, exists only as a maintained graft derived from a single individual ({Service, 2012 #22;Sherwood, 2014 #21}). Due to the significance of *Kokia* to Hawaiian forests, diversity in the genus has been evaluated for the purposes of conservation {Sherwood, 2014 #21}{Morden, 2017 #125}. A remarkable amount of diversity within and among species has been detected, particularly given the demographic history of *Kokia*, which includes the original genetic bottleneck associated with dispersal to the Hawaiian Islands, subsequent inter-island dispersals, and the subsequent bottlenecks due to habitat loss and the introduction of competitive and/or damaging alien species {Morden, 2017 #125}{Sherwood, 2014 #21}.

The native region of *Gossypioides*, the sister genus to *Kokia*, is located 15,000 kilometers distant in East Africa and Madagascar. The two species that comprise the genus, *G. kirkii* M. Mast. and *G. brevilanatum* Hoch. (East Africa and Madagascar, respectively), are themselves reproductively isolated and, with *Kokia*, are cytologically distinct from the remainder of the cotton tribe in that they appear to have experienced an aneuploid reduction in chromosome number. Specifically, while most genera in the *Gossypieae* are based on n=13, species in both *Kokia* and *Gossypioides* are n=12, likely representing a chromosome loss or fusion event. The two species of *Gossypioides* also are cytogenetically distinct, with an unusually long chromosome pair in *G. brevilanatum* {Hutchinson, 1937 #25;Hutchinson, 1943 #29}. Hutchinson (1943) notes that successful grafts can be made between *Kokia drynarioides* and *Gossypioides kirkii*, and that their shared chromosomal reduction (n=12)is unique in the tribe.

Despite extensive research on the evolution of *Gossypium*, these sister genera have been understudied, except for their utility as outgroups for cotton phylogenetic and genomic research {Wendel, 2015 #126} and, in the case of *Kokia*, for assessments of current status and diversity {Morden, 2017 #125}. Direct comparisons of the two genera are limited. Estimates of synonymous substitutions for nuclear gene orthologs indicate that the distance between *K. drynarioides* and *G. kirkii* is less than that between basally diverged species in *Gossypium* {Wendel, 2015 #126}*,* i.e., approximately 2% versus 3.6% {Cronn, 2002 #26}{Flagel, 2012 #128}, although these estimates for *Kokia* and *Gossypioides* are based on few genes. Genomic resources for both genera are minimal and access to plant material is limited. With the recent exception of studies on divergence diversity within and among *Kokia* species noted above, much of our knowledge regarding these genera is decades old {Hutchinson, 1947 #23;Seelanan, 1997 #19;Fryxell, 1968 #28}.

The history of these genera, however, is biogeographically intriguing. The current geographic ranges of *Kokia* in the Hawaiian Islands and *Gossypioides* in East Africa-Madagascar, combined with their sister-genus status and divergence time estimates, implies that there has been at least one significant trans-oceanic traversal to the relatively young Hawaiian archipelago, whose modern members began to emerge only about 4-6 mya {Flinders, 2010 #129}, an age on the same order of magnitude as that estimated for the divergence between *Kokia* and *Gossypioides* {Seelanan, 1997 #19}.

Here we apply a whole-genome sequencing strategy to understand the evolution and divergence of these two genera from a genomic perspective. We present a draft assembly of *Kokia drynarioides*, and compare it to a forthcoming reference-quality sequence of *Gossypioides* *kirkii* (Ramaraj et al., unpublished). Through genome sequence comparisons, we derive a more precise estimate of the divergence time between the two genera, and their similarities and differences with respect to their suite of genes and repetitive sequences. As they are closest genera to the cotton genus, this information may prove informative with respect to understanding the evolution and composition of the cotton genome.

**Methods**

*Sequencing and genome assembly of* Kokia drynarioides

DNA was extracted from mature leaves using the Qiagen Plant DNeasy kit (Qiagen). Total genomic DNA was independently sheared via Covaris into two average sizes, i.e., 350bp and 550bp, for Illumina library construction. A single, independent library was constructed from each fragment pool using the Illumina PCR-free library construction kit (Illumina). The 350 bp library was sequenced on a single lane of Illumina HiSeq2000 and the larger, 550bp library was sequenced on two MiSeq flowcells (both at IGBB, Mississippi State University).

The reads were trimmed and filtered with Trimmomatic v0.32 {Bolger, 2014 #35} with the following options: (1) sequence adapter removal, (2) removal of leading and/or trailing bases when the quality score (Q) <28, (3) removal of bases after average Q <28 (8 nt window) or single base quality <10, and (4) removal of reads <85 nt.

RNA was extracted from three biological replicates of 3cm (length) seedling leaves for both species using the Concert Plant RNA Reagent (Invitrogen) according to the manufacturer’s instructions. Illumina libraries were generated for each RNA using the TruSeq RNA Sample Preparation Kit (Illumina) in preparation for paired-end, 150 nt sequencing. Sequencing was completed on the Illumina HiSeqX Ten at BerryGenomics (Beijing). MEGAHIT commit:02102e1 {Li, 2015 #107} was used to assemble the RNA data into transcripts.

The trimmed DNA data and RNA assembly were jointly assembled via ABySS v2.0.1 {Simpson, 2009 #103}, using every 5th kmer value from 65 through 200. The assembly with the highest E-size {Salzberg, 2012 #104} was retained for improvement and analysis. Each retained assembly was further scaffolded with ABySS using the MEGAHIT-derived transcripts. ABySS Sealer v2.0.1 \cite{Paulino2015} was used to fill gaps in the scaffolded assembly using every 10th kmer starting at 100 and decreasing to 30. Pilon v1.22 {Walker, 2014 #105} polished the resulting gap-filled assembly using all trimmed DNA data. QUAST v4.5 {Gurevich, 2013 #106} was used to generate the final assembly statistics.

*Genome annotation*

MAKER (v2.31.6) {Holt, 2011 #98} annotation of the genome was completed in two rounds, using only contigs >1 kb in size and training MAKER with *Kokia*-specific sequences. First pass *de novo* annotations were derived from Genemark (v4.3.3) {Lomsadze, 2005 #102} and retained for MAKER training. At the same time, BUSCO (v2) BUSCO (v2) {Simão, 2015 #96} was used both to train Augustus and create a Snap model {Korf, 2004 #97}. Finally, Trinity (v2.2.0) {Grabherr, 2011 #99} was used to create an RNASeq-assembly to pass to MAKER as EST evidence. The first pass of MAKER was run using the combination of: (1) the output from Genemark, (2) the BUSCO-generated Snap model, (3) the BUSCO-trained Augustus {Stanke, 2008 #100} model, (4) the Trinity RNASeq-assembly as ESTs, and (5) the UniProt protein database {The UniProt Consortium, 2017 #101}.

After the first pass of MAKER was complete, the annotations generated by MAKER were passed to autoAug.pl (an annotation training script included with Augustus), and were additionally used to generate a second Snap model. MAKER was run again with the same input except using the newly generated Snap model (#2 above) and Augustus model (#3 above) to replace those in the first pass. All annotations were output to gff format and can be found at https://github.com/Wendellab/KokiaKirkii.

*dN/dS Estimation*

Amino acid sequences from *G. kirkii,* *Gossypium raimondii* {Paterson, 2012 #42}, and *K. drynarioides* were clustered using OrthoFinder v1.1.4 {Emms, 2015 #33}, which utilizes a Markov clustering algorithm of normalized BLASTp scores to infer homology between proteins sequences from different species. Default values were used for the inflation parameter (1.5) in the Markov clustering, and the “–og” flag was used to prevent downstream analyses after the orthologous groups were generated. Orthologous groups containing only a single representative from all species were retained and these groups were subsequently filtered if one or more representatives contained ambiguous nucleotide bases (indicating poor sequence coverage). Amino acid sequences from each possible pairwise group (*G. raimondii* + *G. kirkii*, *G. raimondii* + *K. drynarioides*, *K. drynarioides* + *G. kirkii*) were aligned using the pairwise2 python package (https://github.com/biopython/biopython/blob/master/Bio/pairwise2.py) and the BLOSUM62 substitution matrix {Eddy, 2004 #130}; the highest scoring alignment then served as a guide for codon-aligning the CDS sequences using a custom python script (<https://github.com/Wendellab/KokiaKirkii>).

Pairwise *dN* and *dS* values were calculated via CODEML (PAML v.4.9; {Yang, 2007 #31}) and groups with any pairwise *dS* > 0.6 were removed due to possible inclusion of non-orthologous proteins; this threshold represents the upper-limit average of dS values between *G. raimondii* and *Theobroma cacao*, a more distant relative. Distributions of all pairwise *dN, dS,* and *dN/dS* values were evaluated, and basic statistics (mean, median, and standard deviation) were calculated in R (CITATION).

*Estimating Divergence Time*

While many previous estimates of divergence time within Gossypieae {Cronn, 2002 #26} have relied on the estimate of a single gene (AdhA) in Brassicaceae {Koch, 2000 #30} or palms {Morton, 1996 #32}, we sought to utilize the breadth of genomic data available, as well as fossil estimates for the emergence of the Malvaceae (CITATION) to gain a better estimate of the divergence times of members of the Malvaceae. We used the above method for calculating dS values between *Theobroma cacao* and *G. raimondii* but removed any dS values greater than 3 to eliminate saturated synonymous sites (n=43).We then used the equation r = dS/(2T) where T is the fossil-calibrated estimate for the age of the Malvaceae (60 MYA; CITATION), r is the number of synonymous substitutions x synonymous site-1 x year-1 in the Malvaceae, and dS the median of the dS distribution using every identified single copy ortholog. Divergence time between each pairwise group within Gossypieae was estimated using the equation T=*dS*/(2r) where r is the synonymous substitution rate calculated above and dS is the median dS value in the dS distribution for each pairwise comparison (after applying the filtering criteria).

*Copy Number Variation Estimation*

A custom Python script (<https://github.com/Wendellab/KokiaKirkii>) was used to calculate lineage-specific gene losses and duplications between *G. kirkii* and *K. drynarioides*, as inferred by OrthoFinder. First, orthologous groups were filtered for clusters with both copy number variation (CNV) among species and where either *G. kirkii* or *K. drynarioides* had the same copy number as *G. raimondii*. Gene gain or loss was inferred when the non-equal species contained more or fewer genes, respectively, than the species equivalent in copy number to *G. raimondii*. Although an absolute limit on CNV size was not set, most orthologous groups did not have a CNV > 3 genes.Verification of inferred gains and losses was completed by searching for the “missing” genes via gmap {Wu, 2005 #131} of the coding sequence to a masked genome, where all annotated genes are masked.

*Repeat clustering and annotation*

All reads from one of the paired-end files (i.e., R1) were filtered for quality and trimmed to a standard 95nt using Trimmomatic version 0.33 {Bolger, 2014 #35} as per (<https://github.com/Wendellab/KokiaKirkii>). Surviving reads were randomly subsampled to represent a 1% genome size equivalent for each genome {Hendrix, 2005 #36;Wendel, 2002 #37} and combined as input into the RepeatExplorer pipeline {Novák, 2013 #38;Novák, 2010 #39}, which is designed to cluster reads based on similarity and identify putative repetitive sequences using low-coverage, small read sequencing. Clusters containing a minimum of 0.01% of the total input sequences (i.e., 201 reads from a total input of 2,013,469 reads) were annotated by the RepeatExplorer implementation of RepeatMasker {Smit, 2013-2015 #40} using a custom library derived from a combination of Repbase version WHATEVER {Bao, 2015 #41} and previously annotated cotton repeats {Paterson, 2012 #42;Grover, 2008 #58;Grover, 2007 #62;Grover, 2004 #63;Hawkins, 2006 #65}. A cutoff of 0.01% read representation is common; however, we evaluated the suitability of this cut using a log of diminishing returns (FIGURE Cutoff; <https://github.com/Wendellab/KokiaKirkii>).

Within the annotated clusters, the number of megabases (Mb) attributable to that cluster (i.e., element type) for each genome/accession was calculated based on the 1% genome representation of the sample and the standardized read length of 95 nt; total repetitive amounts for each broad repetitive classification were summed from these results. The genome occupation of each cluster (i.e., the calculated number of Mb) was normalized by genome size for each accession, resulting in the percent of each genome occupied by that element type, for use in multivariate visualization (i.e., Principle Coordinate Analysis and Principal Component Analysis). Raw counts were also log-transformed and visualized via PCoA. All analyses were conducted in R {Team, 2017 #66}; R versions and scripts are available at (<https://github.com/Wendellab/KokiaKirkii>).

*Repeat heterogeneity and relative age*

Relative cluster age was approximated using the among-read divergence profile of each cluster, as previously used for *Fritillaria* {Kelly, 2015 #68} and dandelion {Ferreira de Carvalho, 2016 #67}. Briefly, an all-versus-all BLASTn {Boratyn, 2013 #69} {Altschul, 1990 #70} was conducted on a cluster-by-cluster basis using the same BLAST parameters implemented in RepeatExplorer. A histogram of pairwise percent identity was generated for each cluster and the trend (i.e., biased toward high-identity, “young” or lower-identity, “older” element reads) was described for each via regression models using R. Specifically, two regression models were used to describe the data as either linear (Y = a + bX) or quadratic (Y = a + bX + cX2), and the model with the highest confidence was determined via Bayesian Information Criterion {Schwarz, 1978 #71}. The read similarity profile for each cluster was automatically evaluated for each histogram to determine if the reads trend toward highly similar “young” or more divergent “older” reads, as per (Julie paper) with an additional category. These categories include (1) positive linear regression; (2) absence of linear regression; (3) negative linear regression; (4) positive quadratic vertical parabola, trend described by right-side of vertex; (4b) positive quadratic vertical parabola, trend described by left-side of vertex; (5) negative quadratic vertical parabola, trend described by right-side of vertex; and (6) negative quadratic vertical parabola, trend described by left-side of vertex and vertex at >99% pairwise-identity (Figure Age Examples). Categories which trend toward highly identical reads (i.e., 1, 4, and 6) were interpreted as having relatively young membership, whereas categories which trend toward lower identity (i.e., 2, 3, 4b, and 5) were interpreted as being composed of older elements. As with Ferreira de Carvalho (2016), this regression simply provides a relative characterization of cluster/element age and is not designed to detect statistically significant differences.

*Repetitive profiles between* Kokia drynarioides *and* Gossypioides kirkii

Comparison of abundance for the annotated clusters in *Kokia drynarioides* and *Gossypioides kirkii* were visualized via ggplot {Wickham, 2016 #72}, including a 1:1 ratio line to indicate the expected relationship between *K. drynarioides* and *G. kirkii* cluster sizes if their repetitive profiles had remained static post-divergence. Differential abundance (in read counts) between *K. drynarioides* and *G. kirkii* for each cluster was evaluated via two-sample chi2 tests; all p-values were subject to Benjamini-Hochberg correction for multiple testing {Benjamini, 2001 #73}.

*Indel characterization in* Kokia drynarioides *and* Gossypioides kirkii

Indels in *K. drynarioides* and *G. kirkii* were evaluated by mapping each set of sequencing reads to the *G. raimondii* genome and using GATK (v 3.6) {McKenna, 2010 #122;DePristo, 2011 #123;Van der Auwera, 2002 #124} to align and characterize indels. GATK indel calls were pruned to remove (1) positions with missing data in *G. kirkii* or *K. drynarioides* or (2) heterozygous sites. The resulting table was imported into R {Team, 2017 #66} for characterization of indels and length determination using the *G. raimondii* reference state as an outgroup. Indels were characterized as insertions and deletions for each species under the following criteria: (1) the state must be different in *K. drynarioides* and *G. kirkii*; (2) either *K. drynarioides* or *G. kirkii* must share the state with the outgroup; (3) insertions are represented by longer sequence in either *K. drynarioides* or *G. kirkii* compared to the other two; and (4) deletions are represented by shorter sequence in *K. drynarioides* or *G. kirkii* as compared to the other two. Software versions and scripts are available at (<https://github.com/Wendellab/KokiaKirkii>).

**Results**

*Kokia genome assembly and annotation*

ABySS assembly of the 80X coverage Illumina (trimmed; raw = 111X) led to 19,146 scaffolds (25,827 contigs) ranging in size from 500bp to 2.29Mb and comprising a total length of 520.9 Mb (Table XX Assembly Stats; estimated genome size for *K. drynarioides* = 590 Mb {Wendel, 2002 #37}). Nearly 80% of the *K. drynarioides* assembly is represented in scaffolds of >50kb, which, in conjunction with an N50 of 176.7 kb, indicates a relatively contiguous genome. As an additional measure of genic completeness, we searched for 1,440 Benchmarking Universal Single-Copy Ortholog (BUSCO) groups {Simão, 2015 #96} in the *K. drynarioides* assembly. This search recovered 1,377 BUSCOs (95.6%), with 1,213 (84.2%) recovered as single-copy (Table BUSCO). Annotation of the *K. drynarioides* genome (Table Annotation Statistics) resulted in 29,231 gene models, approximately 22% fewer than in the “gold-standard” *Gossypium raimondii* genome sequence {Paterson, 2012 #42}, which has 37,505 predicted protein-coding genes.

For comparative purposes, we annotated the forthcoming *G. kirkii* genome in the same manner as the *K. drynarioides* genome using two iterations of MAKER and *G. kirkii* leaf RNA-seq (Ramaraj, unpublished). The preliminary version of the *G. kirkii* genome used here has greater contiguity than *K. drynarioides*, i.e., an N50 of 616 kb and a total contig length of ~530 Mb; however, BUSCO analysis recovered approximately the same number of complete and single-copy complete BUSCOs (1,349 and 1,213, respectively). The same annotation method also yielded approximately the same number of gene models in *G. kirkii* as in *K. drynarioides* (29,179 versus 29,231).

*Molecular evolution* *between* Kokia drynarioides *and* Gossypioides kirkii

OrthoFinder-based clustering resulted in 21,414 orthologous groups, of which 12,281 contained only one gene from each species (i.e., singleton groups). A disproportionate number of *G. raimondii* genes were not included in any group, as compared to the sister genera (10,408 in *G. raimondii* versus 5,188 and 4,400 in *G. kirkii* and *K. drynarioides*, respectively), an observation consistent with the observation of nearly 8,000 additional gene models in the *G. raimondii* genome. Rates of molecular evolution among these three lineages were estimated for each singleton group, with the exception of those (n=106) where any pairwise comparison resulted in dS > 0.6 (i.e., the upper-estimate of the dS between *G. raimondii* and *T. cacao;* CITATION1). The median dS value for *G. kirkii* vs *K. drynarioides* was approximately half that of either *G. raimondii* vs. *G. kirkii* or *G. raimondii* vs *K. drynarioides* (0.0383 versus 0.0743 and 0.0810 substitutions x synonymous site-1 x yr-1, respectively; Table dN/dS Summary), whose median dS values were approximately equivalent (STAT TEST HERE?). The median dN values for each comparison showed a similar pattern, i.e., 0.0050 between the sister genera versus 0.0086 and 0.0095 substitutions x nonsynonymous site-1 x yr-1 for *G. raimondii* vs. *G. kirkii* and *G. raimondii* vs *K. drynarioides,* respectively (Table dN/dS Summary).

*Divergence Time within Malvaceae*

Clustering of T. cacao and G. raimondii resulted in 13,643 single copy orthologs with a dS value below our threshold (to eliminate saturated synonymous sites). The median of the resultant dS distribution (Figure Malvaceae\_Divergence) was 0.4332, which predicts a synonymous substitution rate (r) of 3.61x10-9 synonymous substitutions x synonymous site-1 x year-1. Using this evolutionary rate estimates that *Gossypium* diverged from the *Kokia* and *Gossypioides* lineage between 10.29 and 11.22 MYA, and the sister genera *Kokia* and *Gossypioides* subsequently diverged from each other approximately 5.30 MYA.

*Copy Number Variation between* Kokia drynarioides and Gossypium kirkii

The 9,133 orthologous groups not classified as singleton groups were evaluated for evidence of CNV (see methods), resulting in 2,991 candidate groups with possible copy number alterations in *G. kirkii* and 2,424 candidates in *K. drynarioides*. The remaining 3,718 groups were excluded either due to complexity (i.e., different copy numbers in each species) or because they were indicative of CNV between *G. raimondii* and *G. kirkii*/*K. drynarioides*, but not between the sister genera themselves..

Candidate CNV groups were evaluated for direction (gain versus loss) and magnitude, i.e., how many genes were gained/lost. We infer 731 genes gained and 2,957 lost in *G. kirkii* (distributed among 259 and 2,730 orthologous groups, respectively; Table CNV\_table). The CNV magnitude (i.e., the number of genes gained or loss per group) varied between one and seven, although two groups encompassed a remarkably large number of genes (i.e., 14 and 225; Table CNV\_table); these were excluded from subsequent calculations as putative falsely annotated transposable elements or errors in the clustering algorithm. In *K. drynarioides*, we infer a somewhat similar number of gains and losses, with 790 genes gained in 499 orthologous groups and 2,008 genes lost from 1,925 orthologous groups. The magnitude of gains varied from one to eight copies, while the magnitude of losses was slightly lower at one to six copies per group (Table CNV\_table). Interestingly, the number of groups where genes were gained in duplicate for *K. drynarioides* (i.e., two genes gained in the same orthologous group) was nearly as high as the number where only one copy was gained (200 vs 260 groups, respectively).

Because overlooked annotations affect our ability to infer CNV events, we evaluated each genome for a subset of the “missing” annotations using only the easiest to interpret cases (i.e., one gene in *G. raimondii* versus >1 or 0 in either *G. kirkii* or *K. drynarioides* for gains and losses, respectively). For the 211 gain events in *G. kirkii* and 394 in *K. drynarioides* evaluated*,* few genes (1 - 6 %) were recovered from the gene-masked genome sequences (see methods), and in most cases, the predicted protein sequence was non-viable (Table CNV\_recovered). For the 2,144 losses in *G. kirkii*, 1,465 were recovered in the masked *G. kirkii*; however, 477 contained frame-shift mutations resulting in non-viable proteins. Likewise, 872 of the 1,458 putative gene losses in *K. drynarioides* found in the non-annotated regions of the *K. drynarioides* genome, with 358 non-viable protein models. Notably, 46% of the *G. kirkii* and 35% of the *K. drynarioides* missing proteins were recovered from the unannotated regions, indicating either missed annotations or deleted genes whose basic sequence remains detectable by the alignment methods used here (i.e., recent deletions). Even if these were all overlooked gene annotations, the number of losses in both species would greatly exceed the number of gains (by about 3-5x); however, the rate of gene loss would become approximately similar (1,156 and 944 losses in *G. kirkii* and *K. drynarioides*, respectively).

*Changes in the repetitive landscape between* Kokia drynarioides *and* Gossypioides kirkii

Because *K. drynarioides* and *G. kirkii* have relatively compact genomes, multiple representatives of three cotton species previously used for repetitive analysis {Renny-Byfield, 2016 #74} were included in the clustering to aid in the identification of repeat-derived sequences. Just over two million reads derived from these five species (comprising 1% genome size equivalents each) were co-clustered using the RepeatExplorer pipeline, producing a total of 74,001 clusters (n >2 reads). Because the smallest clusters are not informative with respect to repetitive sequence evolution, we chose to annotate only those clusters comprising greater than 0.01% of the total reads input (=201 reads); this procedure resulted in 274 retained clusters. We evaluated the cumulative read sum as the cluster number increases (clusters are numbered from largest to smallest) to confirm that this represents a reasonable partitioning of the data set, i.e., most of the input data was represented in the analyzed clusters (cotton\_cutoff.png).

Despite similarly sized genomes, *K. drynarioides* and *G. kirkii* show an approximately 1 Mb difference in clustered repeats (109.4 Mb vs 110.3 Mb, respectively), although this difference is not statistically significant (χ2 p > 0.95). Contingency table analysis of the repetitive profiles of each species, as well as the total amount of repetitive DNA calculated for each, suggest that these profiles are indistinguishable (at p < 0.05), despite being an intergeneric comparison. Interspecies (intragenus) repetitive profiles for *Gossypium* species present in the analysis showed a different pattern, as expected from the two-fold difference in genome size, whereby *G. raimondii* shows a highly distinct repetitive profile (p <0.05) compared to either A-genome species (i.e., *G. herbaceum* and *G. arboreum*). Notably, the two A-genome species are not distinct (see discussion).

To explore further the similarities and differences between the repetitive fractions of the *K. drynarioides* and *G. kirkii* genomes, we considered the possibility that while the overall repetitive profiles may not be significantly different, individual clusters may be. Toward this end, we conducted a χ2 test of independence for each cluster and applied a Benjamini-Hochberg correction for multiple testing. At p<0.05, 55 clusters (out of 188) are differentially abundant in *K. drynarioides* versus *G. kirkii*, with the species displaying greater abundance occurring more frequently in *K. drynarioides* versus *G. kirkii* (34 versus 21 clusters), although the total number of reads in differentially abundant *G. kirkii* clusters was marginally greater (7413 reads versus 7252, representing a 1.5 Mb genome-wide difference). Because these differentially abundant clusters could represent differences in either proliferation or decay/removal, we gauged the relative age of each cluster based of the method of Ferreira de Carvalho et al. (2016). This analysis attempts to characterize the age of each cluster based on the distinctiveness of the reads which comprise the cluster; that is, younger clusters will have reads that are skewed toward high similarity, whereas reads comprising older clusters will have more inter-read differences. While an imperfect measure, this characterization permits a generalized perspective on the repeats identified here. Overall, most of the repeats in *K. drynarioides* and *G. kirkii* displayed a pattern suggestive of older elements (202 “older” versus 72 “young”); however, of the 55 differentially abundant clusters, nearly half (25) were categorized as “younger” (Table Ages). Interestingly, over 80% of the “young” clusters were over-represented in *K. drynarioides*, potentially reflecting differential amplification in these two species.

Most of the clusters were broadly annotated as belonging to the *Ty3/gypsy* superfamily, a result commonly observed in plant genomes (Figure Amounts; {Lee, 2014 #133}{Hawkins, 2006 #65}{Schnable, 2009 #134}{Baucom, 2009 #135}{Tian, 2009 #136}{Paterson, 2009 #137). Overall, gypsy elements comprise 77.6 and 76 Mb of the *K. drynarioides* and *G. kirkii* genomes, respectively, with uncategorized LTR-retrotransposons and *Ty1/copia* elements comprising the next most abundant repeats and in similar amounts in each genome (Table Abundance). Unsurprisingly, the small genomes of *K. drynarioides* and *G. kirkii* had lower absolute abundance of most repeat types than the larger *Gossypium* genomes *except* the predicted non-LTR retrotransposon category, in which these two species had comparable or slightly greater occupation as the cotton species, which possess 2-3x larger genomes (Figure Amounts). This difference is due to the sole retroposon cluster recovered, which was in the top five largest clusters for both *K. drynarioides* and *G. kirkii*. The high percent identity among reads for this cluster suggests it is relatively young, and it has likely experienced recent proliferation in these both species. Furthermore, the cluster shows differential abundance between the two species, suggesting that either the proliferation began prior to species divergence and continued with varying success afterwards, or the two lineages experienced similar releases from repression for this element, although again to varying degrees. The other differentially abundant clusters were largely annotated as putative gypsy elements (61.8 %).

Ancestral state reconstructions for the 22 clusters with the lowest p-value (p<0.001) were conducted using both *K. drynarioides* and *G. kirkii*, as well as three cotton representatives as outgroup species (i.e., *Gossypium raimondii*, *G. arboreum*, and *G. herbaceum*). Patterns of both amplification and deletion were inferred (Figure\_grid.anc.png), sometimes within the same cluster. For example, the repeat represented by cluster 162 has experienced copy number growth in both *K. drynarioides* and *G. kirkii*, with the element attaining much higher copy numbers in *K. drynarioides* (Figure\_grid.anc.png). Likewise, both *K. drynarioides* and *G. kirkii* have experienced reductions in copy number for repeat cluster 5, albeit to different extents. Finally, a large subset of the repeat clusters (20/22) show gain in one of the two lineages coupled with concomitant loss in the other, creating differentially abundant clusters (Figure\_grid.anc.png; see cluster 141 for example). These data implicate a recurring pattern of differential proliferation and removal of multiple different repetitive element families (mostly retrotransposons). Congruent with their equivalent genome sizes, no lineage bias was observed for amplification versus contraction (Figure\_grid.anc.png).

*Patterns of insertion and deletion in* Kokia drynarioides *and* Gossypioides kirkii

To explore further sequence gain and loss in these two genera, we polarized indels (as predicted by GATK; see methods) for both *K. drynarioides* and *G. kirkii* using the *G. raimondii* genome to represent the ancestral state. A gain or loss was inferred when the one taxon shared the reference state with *G. raimondii* and the other had an apparent insertion or deletion. *Kokia drynarioides* exhibited a greater number of both insertions and deletions; that is, of the 490,591 indels that passed our filtering criteria, 130,177 were insertions in *K. drynarioides* and 159,222 were deletions, whereas *G. kirkii* had a total of 87,951 insertions and 113,241 deletions. The distribution of insertion and deletion sizes was biased (for both) towards very small (<10nt) indels; however, when considering the global pattern, insertions in *K. drynarioides* tended to be longer than in *G. kirkii*, whereas *G. kirkii* had a greater number of smaller insertions (Figure\_indels.png). For deletions, *K. drynarioides* and *G. kirkii* were largely similar in the number of smaller deletions; however, *K. drynarioides* exhibited more deletions as the size increased. The overall consequence of these differences in indel evolution resulted in a net gain of 68.6 kb for *K. drynarioides* and a net loss of 113.2 kb in *G. kirkii*, a total genome size difference of ~181.8 kb (0.03% of genome size). The distribution of insertions and deletions across each chromosome was roughly even for both taxa, with up to a two-fold difference in indel number across chromosomes (Figure\_circos.png).

**Discussion**

Divergence and speciation are expected outcomes of long-distance insular dispersal, whose conceptual foundations are rooted in the observations of Darwin and other early evolutionary biologists. The tribe *Gossypieae* is characterized by such dispersals, ultimately achieving worldwide distribution on all tropical and subtropical-inclusive continents. Most *Gossypieae* genera, save for the eponymous *Gossypium* (cotton genus), have been grossly understudied except as each pertains to the evolution of cotton. Here we present a preliminary comparative analysis for the outgroup genera to *Gossypium*, which together provide insight into the interesting biogeographic history of these genera and their equivocality as outgroups in studying the evolution of the cotton genus.

Interest in the sister genera of *Kokia* and *Gossypioides* stems largely from their close evolutionary relationship to the cotton genus (although *Kokia* is an important member of Hawaiian forest communities; see introduction). Initial divergence estimates placed the most recent common ancestor of *Gossypium* and *Gossypioides*/*Kokia* at approximately 10-15 million years before present (MYBP), and the *Kokia* versus *Gossypioides* split in the Pliocene at approximately 3-5 MYA2 (citation, citation). These initial estimates used 12 loci (11 nuclear) to provide the foundation for cotton evolutionary research over the past several decades; however, the ability to address equivalence in evolutionary rates between the two has been lacking until recently.

Here we present a robust, fossil-calibrated estimate for the synonymous substitution rate (3.91x10-9) within Malvaceae using 13,643 single copy orthologs between *G. raimondii* and *T. cacao*. We use this estimate, along with 12,175 nuclear genes within three members of *Gossypieae*, to confirm that the synonymous substitution rate is roughly equivalent between *G. raimondii* and either *G. kirkii* or *K. drynarioides*, suggesting that, despite their disjunct geographic distribution and the known significant founder effect on the Hawaiian *Kokia* species, either (1) there are no large differences in generation time and/or mutation rate per generation between *G. kirkii* and *K. drynarioides*, or (2) that any differences are reciprocal in their effect on the synonymous substitution rate. In contrast to previous analyses, however, which estimated an approximately four-fold difference in divergence time between *Gossypium* and *Gossypioides*/*Kokia,* we estimate only a two-fold difference; that is, the divergence of *Gossypium* from the *Kokia*/*Gossypioides* ancestor occurred approximately twice as long ago as the divergence of those two sister genera. Our estimate of 10.29-11.22 MYA between *G. raimondii* and *G. kirkii*/*K. drynarioides* is similar to previous estimates; however, our estimate that *K. drynarioides*, the Hawaiian endemic, diverged from *G. kirkii* approximately 5.30 MYA, is not only older than the previous range, but is also well before the emergence of the Hawaiian Islands (ca. 3 MYA). Because a signature trait of the Gossypieae tribe is multiple trans-oceanic dispersals, it is not unreasonable to surmise that the evolutionary history of *Kokia* may include multiple trans-oceanic voyages before its arrival in the Hawaiian Islands while undergoing local extinction at any intermediary locales thereby restricting its current ecological distribution to Hawaii.

Regardless of our accuracy with respect to absolute time, it is clear that the divergence between *Gossypium* and that of *Kokia* from *Gossypioides* is closer than previously reported. While indicating a longer time of independent evolution for each sister genus, this is of little consequence in selecting an outgroup for *Gossypium*. Furthermore, despite the demographic history of *Kokia* that would result in the dual pressures of founder-effect and genetic drift (perhaps in multiple rounds), the rate of molecular evolution in these sister genera is surprisingly similar. It bears noting that the demographic history of *Gossypioides* is unknown, and therefore may be quite similar or quite different from that of *Kokia*; regardless, these species remain similar in their genic complement despite pressures from demography.

Repetitive elements, however, are both more labile in nature and potentially sensitive to population size, due to reduced efficiency of purifying selection in populations subject to strong genetic drift {Lefébure, 2017 #75;Lynch, 2003 #76;Yi, 2005 #77;Lynch, 2011 #94;Lynch, 2011 #95}. In the context of genome size, strong drift should lead toward an overall increase in genome size as eukaryotic mutation patterns are typically biased toward insertions, although research addressing the validity and ubiquity of this hypothesis is both scant and conflicting {Whitney, 2011 #79;Gregory, 2008 #81;Whitney, 2010 #78;Arnqvist, 2015 #84;Mohlhenrich, 2016 #93;Lefébure, 2017 #75;Yi, 2005 #77}. While we do not have measures of population size in the present, from the demographic history of *Kokia* *drynarioides* we are aware of a minimum of two severe population reductions: (1) dispersal/speciation on the Hawaiian Islands and (2) dramatic population reduction in response to deforestation. The demographic history of *Gossypioides kirkii* is less clear; the current distribution could also reflect a dispersal event to East Africa, as the ancestral range for the ancestor to these genera is unknown, and the fluctuation in population size for this species is not known. Regardless, given the small current population sizes for both and the population bottlenecks that have affected *Kokia* (minimally), the invariant nature of both their genome size and composition is perhaps surprising. Both species have an estimated genome size of 590 Mb {Wendel, 2002 #37}, which could represent genome size stasis or strong constraints on genome size. Analysis of their global repetitive content suggests that there is only an approximately 1 Mb difference in total (identifiable) repeat content, with very similar profiles for each broad element class. Although these two species are distinguishable via ordination, likely due to the differential abundance of 20% of clusters (55), neither contingency analysis nor Procrustes ANOVA can corroborate this distinctiveness.

Nevertheless, it is clear that differences exist between the two species that likely reflect both gain and loss of sequence. Most of the “younger” differentially abundant clusters that distinguish *K. drynarioides* and *G. kirkii* are over-represented in *K. drynarioides*, a result consistent with the observation that a reduction in population size and concomitant increase in the severity of genetic drift can lead to an increase in insertional mutations (here, 13.7 Mb in *K. drynarioides* versus 10 Mb in *G. kirkii*), possibly due to activation of TEs under stress conditions {Grandbastien, 2004 #108;Kalendar, 2000 #114;Parisod, 2010 #119;Liu, 2003 #120}. Ancestral state reconstructions of TE amounts (Figure\_grid.anc.png) also suggest both gain and loss in *K. drynarioides* and *G. kirkii* (as compared to each other and to *Gossypium*). For the 20 clusters that show gain in one lineage and concomitant loss in the other, gain in *K*. *drynarioides* was inferred for half of the repeats, again including most of the “younger” clusters. These ancestral state reconstructions reflect both sequence gains and losses, better accounting for the static genome size of these species in the face of a changing TE landscape. That is, while the relative age of reads can predict large jumps in occupation of a given repeat, ancestral state reconstructions consider both steady/small transpositional gains relative and loss (although it bears noting that large transpositional bursts will violate the assumption of Brownian motion).

Global patterns of indel formation further extend our understanding of sequence gain and loss beyond that observed in the repeats by providing a genome-wide view agnostic of sequence type. Again, despite their identical genome sizes, *K. drynarioides* and *G. kirkii* vary in their rate of indel formation. In general, *K. drynarioides* insertions are more frequent (130,177 versus 87,951) and longer (mean = 7.4 nt in *K. drynarioides* and 5.8 nt in *G. kirkii*). For deletions, the average size is nearly identical (i.e., approximately 4.8 nt in both); however, the number of deletions in *K. drynarioides* is greater than in *G. kirkii* (159,222 versus 113,241, respectively). These small biases lead to overall gain in sequence for *K. drynarioides* and loss for *G. kirkii* (68.6 kb versus -113.2 kb, respectively), further exaggerating the gain experienced by *K. drynarioides* attributable to “younger” transposable elements (i.e., recent proliferation). These differences also neatly explain why *K. drynarioides* has more “young” TEs whereas *G. kirkii* has more repetitive sequence overall, i.e., the greater deletion rate in *K. drynarioides* is likely contributing to accelerated decay in that lineage.

**Acknowledgements**

The authors acknowledge computational support and assistance from the Iowa State University ResearchIT Unit ([http://researchit.las.iastate.edu/](http://researchit.las.iastate.edu/" \t "_blank)).

Figure Cotton Cutoff : Cumulative sum of the number of reads included in the clusters. The cumulative sum graph displays the percent of reads (y-axis) included in the data analysis given a cluster cutoff (x-axis). The yellow vertical line placed at cluster 274 represents the last cluster containing at least 0.01% of the input dataset.

Figure TEages : Example graphs for regression analyses used for approximate dating. A histogram for percent identity (x-axis) among reads was generated and described via regression models (line), testing both linear (Y = a + bX) and quadratic (Y = a + bX + cX2) models. Five exemplary regression models are shown, including (A) positive linear regression, category 1; (B) negative linear regression, category 3; (C) positive quadratic vertical parabola, trend described by right-side of vertex, category 4; (D) positive quadratic vertical parabola, trend described by left-side of vertex, category 4b; (E) negative quadratic vertical parabola, trend described by right-side of vertex, category 5. Categories 2 and 6 (see methods and Ferreira de Carvalho (2016)) were not observed in this data. Categories 1 and 4 trend toward highly identical reads, indicating the cluster is composed of relatively young elements, whereas categories 3, 4b, and 5 trend toward lower identity, indicative of older (less identical) elements.

Figure TEamounts : The (average) aggregate number of kilobases represented by each transposable element category for each species. Transposable elements were broadly categorized into categories and their representation per species summarized (displayed as a log-scale due to the extensive range in genome occupation among elements).

Figure grid\_anc: Ancestral state reconstruction for the gain/loss of sequence in 22 clusters with the lowest p-value (p<0.001) during the evolution of *Kokia*/*Gossypioides*/*Gossypium*. Total amount of sequence attributable to each cluster is given in kilobases, both next to the name (terminus) and at branch points. Patterns of both amplification (represented by green/blue color) and deletion (yellow/orange/red) were inferred, frequently within the same cluster and sometimes between sister taxa.

Figure indels: The frequency of indels present between *K. drynarioides* (green) and *G. kirkii* (blue), parsed as insertions (top) and deletions (bottom).

Figure circos: Genomic distribution of copy number variations and indels in *K. drynarioides* (Left) and *G. kirkii* (Right).**Outer Ring**: gene gains (dark) and losses (light). **Middle Ring**: insertions*.* **Inner Ring**: deletions.

Figure Malvaceae Divergence: Distribution of synonymous substitution rates (dS) between 13,643 single copy orthologs between *T. cacao* and *G. raimondii*. The median value of the distribution (0.4332) is marked by a vertical black line.