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

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

**Abstract**

**Introduction**

One of the intriguing phenomena that characterizes the cotton tribe, *Gossypieae*, is the prevalence of long-distance, trans-oceanic dispersals*.* The most famous of these occur within the cotton genus itself (*Gossypium*); however, multiple events are found throughout 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}. The sister genera *Kokia* and *Gossypioides* both represent a minimum of one such oceanic dispersal followed by individual regional speciation. Based on molecular divergence estimates derived from both chloroplast and nuclear genes, these genera collectively diverged from the cotton genus during the Miocene approximately 10-15 million years ago (mya; {Seelanan, 1997 #19}{Cronn, 2002 #26}), subsequently splitting into individual genera and achieving widely dispersed, yet very localized ranges.

*Kokia* (Malvaceae) is a small Hawaiian endemic genus composed of four species that were once widespread, major components of Hawaiian forests, yet are now all either endangered, or recently extinct (*K. lanceolata* Lewton; {Bates, 1990 #20}{Sherwood, 2014 #21}). Few individuals remain of the two free-living extant species, *K. kauaiensis* (Rock) Degener & Duvel and *K. drynarioides* (Seem.) Lewton, the latter of which is critically endangered and 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}). The native region of its sister genus, *Gossypioides*, is located over 15,000 kilometers away 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}.

Despite the extensive research on the evolution of *Gossypium*, these sister genera have been grossly understudied, except in serving as phylogenetic outgroups for cotton phylogenetic and genomic research {Seelanan, 1997 #19}{Cronn, 2002 #26}. Genomic resources in both genera are minimal, access to plant material is limited, and with the recent exception of a study by Sherwood and Morden (2014) on diversity among *Kokia* species, 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 intriguing. The current distribution of *Kokia* in the Hawaiian Islands and *Gossypioides* in East Africa-Madagascar necessitates at least one significant trans-oceanic traversal to a relatively young island chain that began to emerge only about 3.4 mya, an age approximately equivalent to the estimated divergence between *Kokia drynarioides* and *Gossypioides kirkii* {Seelanan, 1997 #19} and slightly more recent than the basal most divergence in *Gossypium*. Diversity within *Gossypioides* is unknown, aside from acquisition of reproductive isolation between its sole two species; however, diversity in *Kokia* has been evaluated for the purposes of conservation {Sherwood, 2014 #21}. 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 of the founder, range expansion, and the subsequent bottleneck of habitat loss and the introduction of competitive and/or damaging alien species {Sherwood, 2014 #21}.

Direct comparisons of these genera are limited. Hutchinson (1943) notes that successful grafts can be made between *Kokia drynarioides* and *Gossypioides kirkii*, and their shared chromosomal reduction (n=12)is unique in the tribe. Estimates using a small number of nuclear genes suggest that genic distance between *K. drynarioides* and *G. kirkii* are similar to estimates between basally diverged species in *Gossypium,* i.e., approximately 2% versus 3%, although a slight increase in replacement site substitutions is observed {Cronn, 2002 #26}.

Here we apply a whole-genome sequencing strategy to understanding the evolution and divergence of these two genera, which collectively are the closest relatives of the cotton genus *Gossypium*. We present the first draft assembly of *Kokia drynarioides*, and compare it to the sequence of *Gossypioides* *kirkii* (citation of Gk paper). Through genome sequence comparisons, we derive a precise estimate of the divergence between these two genera, and provide a foundation for a reference sequence to use as a phylogenetic outgroup to *Gossypium*.

**Methods**

*Sequencing and genome assembly*

DNA was extracted from mature leaves using the Qiagen Plant DNeasy kit (Qiagen). 350Bp and 550 bp Illumina PCR Free libraries were made and sequenced on 2 Miseq flowcells and 1 Hiseq 2000 lane at the IGBB. The data were trimmed and filtered with Trimmomatic v0.32 {Bolger, 2014} 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 .... MEGAHIT commit:02102e1 {Li, 2015} was used to assemble the RNA data into transcripts.

The trimmed DNA data and RNA assembly were assembled via ABySS v2.0.1 {Simpson, 2009}, using every 5th kmer value from 65 through 200. The assembly with the highest E-size {Salzberg, 2012} was used in further analyses. Next the selected assembly was further scaffolded with ABySS using the assembled transcripts. ABySS Sealer v2.0.1 {Paulino, 2015} was used to fill gaps in the scaffolded assembly. For each trimmed PCR Free library, Sealer was run with every 10th kmer starting at 100 and decreasing to 30. Pilon v1.22 {Walker, 2014} polished the resulting gap-filled assembly using all the trimmed DNA data.

*Genome annotation*

Several programs were used to generate input for MAKER (v2.31.6) [2] . Trinity (v2.2.0) [1] was used to create an RNASeq-assembly that was passed to MAKER as ESTs. The genome was filtered to remove sequences less than 1kb. With the filtered genome, Genemark (v4.3.3) [3] was used to generate gene predictions and BUSCO (v2) [4] was used to train Augustus and create a Snap model. The first pass of MAKER was run using the output from Genemark, the Snap model created from BUSCO's output, the Augustus [5] model trained by BUSCO, the RNASeq-assembly from Trinity as ESTs, and UniProt as a protein database.

After the first pass of MAKER was complete, the annotations generated by MAKER were passed to autoAug.pl, a script included with Augustus that trains Augustus. These annotations were also used to generate a second Snap model. MAKER was run again, replacing the Snap model and Augustus model from BUSCO with the models generated from the output of the first pass of MAKER.

*Identification of Orthologs*

Amino acid sequences from *G. kirkii,* *G. raimondii* and *K. drynarioides* were clustered using OrthoFinder v1.1.41 {Emms, 2015 #33}, which utilizes a Markov clustering algorithm of normalized BLASTp scores to infer homology between proteins sequences of different species. OrthoFinder is similar to OrthoMCL2 {Li, 2003 #34}, but reduces the number of BLAST results by filtering scores based on reciprocal best hits (RBHs) and corrects for gene length biases and floor-limitation of e-values in BLAST scores prior to clustering. These corrections have been shown to increases precision by improved clustering of singletons (i.e., groups in which only one gene from each species is present) instead of entire gene families into a given orthologous group. 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 groups were generated.

*dN/dS Estimation and Timing of Divergence*

Singletons inferred from OrthoFinder were separated into all 3 possible pairwise groups (Gr + Gk, Gr + Kd, Kd + Gk). Amino acid sequences from each pairwise group were then aligned using the pairwise2 python package and the BLOSUM62 substitution matrix. The highest scoring alignments were then used as a guide to codon-align the CDS sequences. The CODEML package in PAML {Yang, 2007 #31} was used to calculate the dN, dS, and dN/dS values. Singletons in which any pairwise comparison resulted in a dS value greater than 0.03 was removed from the analysis and inferred to be a cluster of non-orthologous proteins. Distributions of all pairwise dN, dS, and dN/dS values were then plotted, and mean value and standard deviation is reported. Estimates of total divergence time between each pairwise group was calculated using the equation T=dS/(2r) where r is the absolute rate of synonymous substitutions of Adh genes in palms (2.6 X 10-9 substitutions X substitution site-1 X year-1) {Cronn, 2002 #26}{Morton, 1996 #32} or members of Brassicaceae (1.5 X 10-8 substitutions X synonymous site-1 X year-1) {Koch, 2000 #30}.

*Copy Number Variation Estimation*

A custom Python script (<https://github.com/Wendellab/KokiaKirkii>) was used to calculate lineage-specific gene losses and duplications as inferred by OrthoFinder. A gene loss was defined as an orthologous group in which 2 species had the same number of genes present (*n*), but the third species contained *n-1* genes. Likewise, a gene duplication was identified by 2 species containing *n* genes, while the third contained *n+1*.

*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 WHATEVER; <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). 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 WHATEVER). 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}.

**Results**

*Kokia genome assembly and annotation*

STATS ON THE KOKIA GENOME HERE. STATS ON THE ANNOTATION TOO.

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

1. Outgroup equivalency/utility: are they equal for molecular evolutionary purposes
   1. Limited by no population data
   2. Ks/Ks of Gk-Gr versus Kd-Gr; are they equivalent
   3. Gene cluster comparisons: does Gk or Kd perform equivalently?
      1. i.e., number of Gr-Kd only groups versus number of Gr-Gk only groups
   4. when would having two outgroups be of a benefit
   5. Ka/Ks for Gk-Kd: high or low? What do we expect?
   6. Gene content comparison : what is “missing”? What is unique?
2. Colinearity (at all?) or just intergenic SNPs/indels via gatk?

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

Because *K. dryanarioides* 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 74,001 clusters (n >2 reads). Because the smallest clusters are neither informative nor reliable indicators of repetitiveness, we chose to annotate only those clusters composed of greater than 0.01% of the total reads input (=201 reads), resulting 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 (cotton\_cutoff.png).

Despite identically sized genomes, *K. drynarioides* and *G. kirkii* show an approximately 1 Mb difference in clustered repeats, although this lacks statistical significance. 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 the intergeneric comparison. Interspecies (intragenus) repetitive profiles for those *Gossypium* species present in the analysis showed a different pattern, whereby the basally divergent *G. raimondii* compared to either A-genome species (i.e., *G. herbaceum* and *G. arboreum*) shows a highly distinct repetitive profile (p <0.05), although, notably, the sister A-genome species are not distinct (see discussion).

To ascertain the extent of the differences between *K. drynarioides* and *G. kirkii*, 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 chi2 test of independence for each cluster and applied a Benjamini-Hochberg correction for multiple testing. At p<0.05, XXX clusters (out of XXX) are differentially abundant in *K. drynarioides* versus *G. kirkii*, with the species displaying greater abundance occurring approximately the same number of times for both (XXX with greater abundance in *K. drynarioides* versus XXX in *G. kirkii*; Table Abundance). 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 (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 highly similar, whereas older clusters will have reads that show a number of 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 versus 72 “young”); however, of the XXX differentially abundant clusters, XXX were categorized as “young” and XXX as “older” (Table Abundance), potentially reflecting SOMETHING ABOUT GAIN VERSUS LOSS.

Most of the clusters were broadly annotated as belonging to the *Ty3/gypsy* superfamily, a result not surprising for a plant lineage (Figure Amounts). Overall, gypsy elements comprise XXX to XXX of the *K. drynarioides* and *G. kirkii* genomes, respectively, with uncategorized LTR-retrotransposons and *Ty1/copia* elements comprising the next most abundant repeats and comprising similar amounts in each genome. Unsurprisingly, the small genomes of *K. drynarioides* and *G. kirkii* had lower absolute abundance of most repeat types *except* the predicted non-LTR retroposons, in which these two species had comparable or slightly greater occupation as the cotton species, which possess 2-3x larger genomes. This difference is due to the sole retroposon clusters recovered, which was in the top 5 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 proliferation in 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 (RIGHT?) (XX %).

**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 first-pass genome assemblies for the outgroup congeners 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.

1. Compare molecular differences to perceived degree of morphological differentiation?

Phylogenetics in the tribe: ndhF shows longer NJ branch length for Kokia than kirkii (congruence and consensus)

Long-distance salt water dispersal common in gossypieae

Advance Agronomy

Lebronnecia – marquesas (south pacific)

Thespecia thespesioides – pan tropical

Hampia – neotropical (americas)

Thespesia populnea – pan tropical

Cephalohibiscus – new guinea and solomon islands (Australia)

Maybe we would expect there to be stepping speciation among these island regions, e.g., south pacific lebronnecia to be between Kokia and kirkii, or neotropical Hampea to be between the two. Clearly congeners, molecularly and united by n=12. Hawaiian islands only ~3myo, so Kokia probably colonized them as they were formed. What about kirkii? Is it an older population, from which Kokia is derived (probably not given the data), or was it a dispersal event from who knows where of a now extinct ancestor?

*Justin’s Bib (Please don’t edit until final)*

1. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol*. 2015;16(1):157. doi:10.1186/s13059-015-0721-2.

2. Li L, Stoeckert CJJ, Roos DS. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes -- Li et al. 13 (9): 2178 -- Genome Research. *Genome Res*. 2003;13(9):2178-2189. doi:10.1101/gr.1224503.candidates.

Tony Bibliography

1. Bolger, A. M.; Lohse, M. & Usadel, B. (2014), 'Trimmomatic: a flexible trimmer for Illumina sequence data', *Bioinformatics* **30**(15), 2114--2120.
2. Salzberg, S. L.; Phillippy, A. M.; Zimin, A.; Puiu, D.; Magoc, T.; Koren, S.; Treangen, T. J.; Schatz, M. C.; Delcher, A. L.; Roberts, M.; Marcais, G.; Pop, M. & Yorke, J. A. (2012), 'GAGE: A critical evaluation of genome assemblies and assembly algorithms', *Genome Research* **22**(3), 557--567.
3. Simpson, J. T.; Wong, K.; Jackman, S. D.; Schein, J. E.; Jones, S. J. M. & Birol, I. (2009), 'ABySS: A parallel assembler for short read sequence data', *Genome Research* **19**(6), 1117--1123.

Adam Bibliography

[1] Grabherr, M., Haas, B., Yassour, M., Levin, J., Thompson, D., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., and Zeng, Q. et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29, 644-652.

[2] Holt, C., and Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12, 491.

[3] Lomsadze, A. (2005). Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Research 33, 6494-6506.

[4] Simão, F., Waterhouse, R., Ioannidis, P., Kriventseva, E., and Zdobnov, E. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210-3212.

[5] Stanke, M., and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19, ii215-ii225.