**Comparative genomics of an unusual biogeographic disjunction in the cotton tribe (Gossypieae) yields insights into genome downsizing**

Corrinne E Grover1, Mark A Arick II2, Justin L Conover1, Adam Thrash2, Guanjing Hu1, William S Sanders2,3,4, Rubab Naqvi5, Muhammad Farooq5, Xiaochong Li6, Lei Gong6, Joann Mudge7, Thiru Ramaraj7, Joshua A Udall8, Daniel G Peterson2, Jodi Scheffler9, Brian Scheffler9, and Jonathan F Wendel1

1 Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA USA

2 Institute for Genomics, Biocomputing, and Biotechnology, Mississippi State University, MS USA

3 Department of Computer Science & Engineering, Mississippi State University, MS, USA

4 The Jackson Laboratory, CT, USA

5 National Institute for Biotechnology and Genetic Engineering (Nibge), Faisalabad, Punjab, Pakistan

6 School of Life Sciences, Northeast Normal University, Changchun, P.R. China

7 National Center for Genome Resources, Santa Fe, NM USA

8 Department of Plant and Wildlife Sciences, Brigham Young University, Provo, UT USA

9 Jamie Whitten Delta States Research Center, USDA-ARS, Stoneville, MS USA

**Abstract**

Long-distance insular dispersal is associated with divergence and speciation because of founder effects and strong genetic drift. The cotton tribe (Gossypieae) has experienced multiple trans-oceanic dispersals, generating an aggregate geographic range that encompasses much of the tropics and subtropics world-wide. Two genera in the Gossypieae, *Kokia* and *Gossypioides*, exhibit a remarkable geographic disjunction, being restricted to the Hawaiian Islands and Madagascar/East Africa, respectively. We assembled *de novo* genome sequences to address questions regarding the age of divergence of these two genera from each other and from their sister-group, *Gossypium*, and processes of genome evolution that have led to genome downsizing. Using 13,000 gene orthologs and synonymous substitution rates, we show that the two disjuncts last shared a common ancestor about 5 MYA, about have as long as their divergence from the cotton genus. We report relative stasis in the transposable element fraction. In comparison to *Gossypium,* there is loss of approximately 30% of the gene content in the two disjunct genera and a history of genome-wide accumulation of deletions. The genomic analyses presented here elucidate genomic consequences of the demographic and biogeographic history of these closest relatives of *Gossypium*, enhancing their value as phylogenetic outgroups.

=^. .^= nice cat! --Malcolm’s contribution

**Introduction**

One of the intriguing evolutionary phenomena that characterizes the cotton tribe, Gossypieae, is the prevalence of long-distance, trans-oceanic dispersal (Wendel and Grover 2015)*.* 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) 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 (Stephens 1958, 1966, Fryxell 1979, Wendel 1989, Wendel and Percival 1990, Wendel and Percy 1990, Dejoode and Wendel 1992, Wendel and Albert 1992, Seelanan et al. 1997, Wendel and Cronn 2003). One example includes the sister genera *Kokia* and *Gossypioides*, from Hawaii and southeast Africa, respectively. Based on preliminary molecular divergence estimates derived from 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 et al. 1997, Cronn et al. 2002).

*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)(Bates 1990, Sherwood and Morden 2014, Morden and Yorkston 2017)). 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, Sherwood and Morden 2014)). Due to the significance of *Kokia* to Hawaiian forests, diversity in the genus has been evaluated for the purposes of conservation (Sherwood and Morden 2014, Morden and Yorkston 2017). A surprising 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 (Sherwood and Morden 2014, Morden and Yorkston 2017).

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 and Ghose 1937, Hutchinson 1943). 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 and Grover 2015) and, in the case of *Kokia*, for assessments of current status and diversity (Morden and Yorkston 2017). 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 and Grover 2015)*,* i.e., approximately 2% versus 3.6% (Cronn et al. 2002, Flagel et al. 2012), 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, Fryxell 1968, Seelanan et al. 1997).

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. The present islands began to emerge only about 4-6 mya (Flinders et al. 2010), an age on the same order of magnitude as that estimated for the divergence between *Kokia* and *Gossypioides* (Seelanan et al. 1997).

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 these species represent the two 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 et al. 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 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 et al. 2015) 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 et al. 2009), using every 5th kmer value from 65 through 200. The assembly with the highest E-size (Salzberg et al. 2012) was retained for improvement and analysis. Each retained assembly was further scaffolded with ABySS using the MEGAHIT-derived transcripts. ABySS Sealer v2.0.1 (Paulino et al. 2015) was used to fill gaps in the scaffolded assembly using every 10th kmer starting at 100 and decreasing to 30. Pilon v1.22 (Walker et al. 2014) polished the resulting gap-filled assembly using all trimmed DNA data. QUAST v4.5 (Gurevich et al. 2013) was used to generate the final assembly statistics.

*Genome annotation*

MAKER (v2.31.6) (Holt and Yandell 2011) 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 et al. 2005) and retained for MAKER training. At the same time, BUSCO (v2) BUSCO (v2) (Simão et al. 2015) was used both to train Augustus and create a Snap model (Korf 2004). Finally, Trinity (v2.2.0) (Grabherr et al. 2011) 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 et al. 2008) model, (4) the Trinity RNASeq-assembly as ESTs, and (5) the UniProt protein database (The UniProt Consortium 2017).

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 et al. 2012), and *K. drynarioides* were clustered using OrthoFinder v1.1.4 (Emms and Kelly 2015), 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); 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)) 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 (Team 2017).

*Estimating Divergence Times*

Earlier estimates of divergence times within the Gossypieae (Cronn et al. 2002) relied on dating calibrations derived from a single nuclear gene (*AdhA*) in the Brassicaceae (Koch et al. 2000) or palms (Morton et al. 1996), and on rates of chloroplast DNA evolution (Seelanan et al., 1999). More recently, divergence times for the Malvaceae have been reported based on a single gene each from the chloroplast and the nuclear genomes (Richardson et al. 2015) suggesting that chocolate (*Theobroma*) and *Gossypium* diverged *circa* 60-70 mya. Using a more extensive data sent, absolute rates of synonymous substitutions have been estimated for eight angiosperm families (De La Torre et al. 2017), fortuitously including the Malvaceae (rate of substitution between *Theobroma* and *Gossypium* estimated to be 4.56E–09/year, based on 42 genes). Here we extended this analysis to include two orders of magnitude more genes (n = 13,643 single copy orthologs) using published genome sequences for these taxa. In estimating dS values between *Theobroma cacao* and *G. raimondii* we removed values greater than 3 to eliminate saturated synonymous sites (43 genes).We then used the equation r = dS/(2T), where T is the fossil (perhaps 60 MYA; (Carvalho et al. 2011) and [www.timetree.org](file:///C:\Users\corrinne\AppData\Local\Microsoft\Windows\INetCache\Content.Outlook\BJV9K5HR\www.timetree.org)) and sequence-calibrated estimate for divergence of *Gossypium* and *Theobroma*, 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 the 13,643 single copy orthologs. 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 and Watanabe 2005) of the coding sequence to a masked genome, where all annotated genes are masked. Genes annotated in *G. raimondii* that were not present in either *K. drynarioides* or *G. kirkii*, i.e., inferred losses, were similarly verified via gmap against the unmasked *K. drynarioides* or *G. kirkii*. Results were visualized with Circos (Krzywinski et al. 2009).

*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 et al. 2014) as per (<https://github.com/Wendellab/KokiaKirkii>). Surviving reads were randomly subsampled to represent a 1% genome size equivalent for each genome (Wendel et al. 2002, Hendrix and Stewart 2005) and combined as input into the RepeatExplorer pipeline (Novák et al. 2010, Novák et al. 2013), 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 et al. 2013-2015) using a custom library derived from a combination of Repbase version 21.08 (Bao et al. 2015) and previously annotated cotton repeats (Grover et al. 2004, Hawkins et al. 2006, Grover et al. 2007, Grover et al. 2008a, Paterson et al. 2012). 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); 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 et al. 2015) and dandelion (Ferreira de Carvalho et al. 2016). Briefly, an all-versus-all BLASTn (Boratyn et al. 2013) (Altschul et al. 1990) 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 using the Bayesian Information Criterion (Schwarz 1978). 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 previously characterized (Ferreira de Carvalho et al. 2016) but 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 that trend toward highly similar reads (i.e., 1, 4, and 6) were interpreted as representing more recent divergences, whereas categories with lower identities (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), 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 and Yekutieli 2001).

*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) (Van der Auwera et al. 2002, McKenna et al. 2010, DePristo et al. 2011) 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) for characterization of indels and length determination using the *G. raimondii* reference state as an outgroup. Indels were characterized as insertions or 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 et al. 2002)). 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 et al. 2015) 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 et al. 2012), 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 (5,982 verified as “missing”; see methods, Table CNV\_recovered). 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*, see methods). 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. 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* gene models resulted in 13,643 single copy orthologs with a dS value below our threshold (to eliminate saturated synonymous sites and likely paralogs). The median of the resultant dS distribution (Figure MalvaceaeDivergence) was 0.4332, which predicts a synonymous substitution rate (r) of 3.61x10-9 synonymous substitutions x synonymous site-1 x year-1, similar to that reported recently for 42 genes (De La Torre et al., 2017). Using this evolutionary rate, we estimate that *Gossypium* diverged from the *Kokia* and *Gossypioides* lineage between 10.29 and 11.22 MYA, and that the sister genera *Kokia* and *Gossypioides* 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 et al. 2016) 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; (Hawkins et al. 2006, Baucom et al. 2009, Schnable et al. 2009, Tian et al. 2009, Lee and Kim 2014){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* for 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 species. Furthermore, the cluster shows differential abundance between the two species, suggesting either that the proliferation began prior to species divergence and continued differentially afterwards, or that 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 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 many subsequent evolutionary biologists. Because of the small population sizes associated with dispersal-mediated genetic bottlenecks, islands serve as natural laboratories to study the effects of isolation and drift on character evolution, including, as we show here, on genome structure and features. The tribe *Gossypieae* is characterized by multiple long-range dispersals, ultimately achieving an aggregate geographic distribution that encompasses tropical and subtropical regions worldwide. With the exception of the type genus *Gossypium*, little is known about the genomes of genera in the *Gossypieae*, apart from estimates of genome size {Wendel, 2002 #37}. Here we present a comparative analysis for the clade of two genera that together comprise the phylogenetic outgroup to *Gossypium*. We provide insight into the interesting biogeographic history of these genera and clarify the temporal framework for divergence between *Gossypioides* and *Kokia* as well as these two genera from *Gossypium*. This framework permits an analysis of the pace, patterns and processes that have characterized genomic divergence among the three genera, including novel insights into gene loss, structural variation, and genome downsizing.

***Temporal framework for divergence and biogeographic implications***. Interest in the sister genera of *Kokia* and *Gossypioides* stems largely from their close evolutionary relationship to *Gossypium*, although *Kokia* is an important member of Hawaiian forest communities (see introduction). Early 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 MYA (Seelanan et al. 1997, Cronn et al. 2002). These initial estimates were from the pre-genomics era, and hence were based on relatively few nuclear and plastid genes. Here we present a robust estimate for the synonymous substitution rate (3.91x10-9 substitutions per site per year) within the Malvaceae using 13,643 single copy orthologs from *G. raimondii* and *T. cacao*. We use this estimate, and a set of 12,175 nuclear orthologs inferred from the three genera of the *Gossypieae*, to confirm that the synonymous substitution rates are equivalent between *G. raimondii* and either *G. kirkii* or *K. drynarioides.* This indicates that despite their disjunct geographic distribution and multiple sequential founder events, there are no significant differences in generation time and/or mutation rate per generation between *G. kirkii* and *K. drynarioides* or that any such differences are reciprocal in their effects. With respect to dating divergences, our genome-scale data set permits us to refine earlier estimates. Thus, in contrast to previous analyses, 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* common ancestor occurred approximately twice as long ago as the divergence of those two sister genera from each other. Our estimate of 10.29-11.22 MYA for the divergence of *Gossypium* from *G. kirkii*/*K. drynarioides* is similar to previous estimates (Seelanan et al. 1997, Cronn et al. 2002, Senchina et al. 2003), which is remarkable observation given the fact that earlier estimates were based on two orders of magnitude fewer genes.

The indication that *K. drynarioides* diverged from *G. kirkii* approximately 5.30 MYA, instead of 3 MYA as reported earlier, may be biogeographically significant in that it suggests a divergence at about the same time as the emergence of the present Hawaiian Islands. Because a signature trait of the Gossypieae is multiple trans-oceanic dispersals, these divergence data suggest that the evolutionary history of *Kokia* (and any now-extinct members of its clade) may have included multiple trans-oceanic voyages before its arrival and evolution in the Hawaiian Islands along with local extinction at geographically intermediate locations. We note, however, that the Hawaiian Islands are the world’s most isolated oceanic archipelago, without clear “stepping stones” across the Pacific Ocean from either continental hemisphere. Alternatively, the antecedent of modern *Kokia* may have made a great leap circa 5.3 million years ago, to a part of the Hawaiian archipelago that presently is eroded and submerged, with subsequent island-hopping as suitable habitat became available during the genesis and ecological development of the island chain. In any event, the biogeographic story is a remarkable one, as the two genera *Kokia* and *Gossypioides* are separated by a minimum of 17,500 kilometers, and yet are each other’s closest relatives. This is even more striking when one considers that present species lack any clear mechanism for oceanic dispersal, as seeds sink relatively quickly. Seeds of many taxa in the tribe do possess a certain degree of salt-water tolerance (Stephens 1958, Fryxell 1979, Wendel and Grover 2015), however, so the possibility remains that this remarkable dispersal voyage entailed some sort of natural rafting on oceanic debris, either of seeds or of mature but undehisced capsules.

***Extensive gene removal differentiates* Kokia *and* Gossypioides.** The temporal framework provided above provides the opportunity to explore the relative evolutionary rates of genomic differentiation. With respect to genes, variation in gene content among species and individuals is more extensive than once thought, leading to the concept of “core” and “dispensable” genomes (together, the pan-genome; (Medini et al. 2005, Hirsch et al. 2014)). Research in plants (Morgante et al. 2007, Springer et al. 2009, Swanson-Wagner et al. 2010, Cao et al. 2011, Chia et al. 2012, Hirsch et al. 2014) suggests that many plant species exhibit evidence of a pan-genome whose “dispensable” component may contribute to diversity and adaptation (Medini et al. 2005, Tettelin et al. 2005, Kahlke et al. 2012). Here, using a divergence time 5.3 million years, we estimate that gene deletions between *Kokia* and *Gossypioides* have occurred at about 188-377 per lineage per million years. Perhaps more surprising is the number of additional genes in the *Gossypium raimondii* genome as compared to that in either *Kokia* or *Gossypioides* (n=~6,000). As gene deletions dramatically outweigh insertions, we infer that many of these represent shared deletions that occurred in the ~5-6 MY between the divergence of *Gossypium* from proto-*Kokia*/*Gossypioides* and the divergence of the two sister genera. This rate of gene deletion would be much higher than in either lineage alone, resulting in approximately ~1,000 deletions per million years in the proto- *Kokia*/*Gossypioides* lineage. Post-divergence, the rate of gene deletion between the two lineages was significantly slower and approaching equivalence (at most a two-fold difference, see results).

***Static genome size in the face of a changing repetitive element landscape.*** Repetitive elements are both labile in nature and potentially sensitive to population size, due to reduced efficiency of purifying selection in small populations because of the prominence of strong genetic drift (Lynch and Conery 2003, Yi and Streelman 2005, Lynch 2011, Lynch et al. 2011, Lefébure et al. 2017). 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 (Yi and Streelman 2005, Gregory and Witt 2008, Whitney and Garland 2010, Whitney et al. 2011, Arnqvist et al. 2015, Mohlhenrich and Mueller 2016, Lefébure et al. 2017). While we do know historical population sizes in the present study, it is clear that population bottlenecks must have been profound in Kokia, as described above. 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 et al. 2002), representing genome size stasis during about 5 million years of divergence. Analysis of their global repetitive content suggests that there is only a trivial (approximately 1 Mb difference) in total (identifiable) repeat content, with very similar overall repetitive profiles for each. We note that this result contrasts with the expectation based on small effective population size alone.

Notwithstanding the relative genomic stasis of the two genera, it is clear that the differences that do exist between the two species reflect both gain and loss of repetitive 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, possibly due to activation of TEs under stress conditions (Kalendar et al. 2000, Liu and Wendel 2003, Grandbastien 2004, Parisod et al. 2010). Ancestral state reconstructions of TE amounts (Figure\_grid.anc.png) also suggest both gain and loss in *K. drynarioides* and *G. kirkii* of approximately the same magnitude, which accounts for the static genome size of these species in the face of a changing TE landscape.

***Rates of indel formations compensate for biased TE proliferation.*** While transposable elements are capable of substantially altering genome size and structure, the presence of indels also contributes to genome size and collinearity (Petrov 2002, Gregory 2003, Vitte and Bennetzen 2006, Hjelmen and Johnston 2017, Kapusta et al. 2017). Previous work in cotton suggests there exist small differences in rates between species with large and small genomes that contribute to overall genome size change (Grover et al. 2008b). Global patterns of indel formation, as inferred from modern sequencing, can further extend our understanding of sequence gain and loss by providing a genome-wide view agnostic of sequence type (e.g., TE-derived) or region. As with the repetitive elements, *K. drynarioides* and *G. kirkii* vary in their rate of indel formation despite their equivalent genome sizes. In general, *K. drynarioides* experiences insertions and deletions more frequently, and the insertions tend to be longer than those found in *G. kirkii* (deletion sizes are equivalent on average). These small biases lead to overall gain in sequence for *K. drynarioides* (+68.6kb) and loss for *G. kirkii* (-113.2 kb), further exaggerating the gain experienced by *K. drynarioides* attributable to “younger” transposable elements (i.e., recent proliferation). In addition, these differences also 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.

***Conclusions***

External influences on genome evolution are many and complex, affecting genomes in sometimes predictable, and sometimes enigmatic, ways. Despite the strong pressures associated with repeated genetic bottlenecks as *Kokia* and *Gossypioides* underwent island dispersal, the most labile component of the genome (i.e., transposable elements) remained surprisingly constant. Furthermore, the changes in size due to differential transposable element occupation were ultimately offset by differential rates of deletion in the two species, resulting in equivocal genome sizes despite ca. 5 MY of independent evolution, strong founder effects, and intense genetic drift. This is perhaps even more remarkable considering that, in approximately the same timeframe (the last 5-10 MY), the related genus *Gossypium* has experienced far more significant changes in genome size due to differential transposable element proliferation, which has led to a 3-fold difference in genome size among cotton species, and similar rates of indel formation (Grover et al. 2008b).

Perhaps more unexpected were the presence of more than 10,000 genes in the *Gossypium raimondii* genome where no *K. drynarioides* or *G. kirkii* homolog was detected, resulting in nearly 8,000 more annotated genes in the *G. raimondii* genome than in either *K. drynarioides* or *G. kirkii*. While some of these additional gene models may be due to differences in annotation methods between *G. raimondii* and *K. drynarioides*/*G. kirkii*, it nevertheless suggests a higher rate of gene deletion in these sister genera. The deletions inferred here, both lineage-specific and those occurring in proto-*Kokia*/*Gossypioides*, are not only interesting from an evolutionary standpoint, but are also germane to the selection of either species as an outgroup to *Gossypium*. While both species can individually serve as useful representatives of the cotton ancestor, it is clear that enough differences exist between the two outgroup genera to warrant inclusion of both as representatives of the ancestral cotton genome.

**Acknowledgements**

The authors acknowledge computational support and assistance from the Iowa State University ResearchIT Unit (<http://researchit.las.iastate.edu/>). We thank the National Science Foundation for support.

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.

**References cited**

Altschul, S F, Gish W, Miller W, Myers E W and Lipman D J (1990) Basic local alignment search tool. Journal of Molecular Biology 215(3): 403-410.

Arnqvist, G, Sayadi A, Immonen E, Hotzy C, Rankin D, Tuda M, Hjelmen C E and Johnston J S (2015) Genome size correlates with reproductive fitness in seed beetles. Proc Biol Sci 282(1815).

Bao, W, Kojima K K and Kohany O (2015) Repbase update, a database of repetitive elements in eukaryotic genomes. Mobile DNA 6(1): 11.

Bates, D M (1990). Malvaceae.In: Wagner W, Herbst D and Sohmer S Manual of the flowering plants of hawai ‘i revised edition. Honolulu, University of Hawai'i and Bishop Museum Press**:** 868-902.

Baucom, R S, Estill J C, Chaparro C, Upshaw N, Jogi A, Deragon J-M, Westerman R P, SanMiguel P J and Bennetzen J L (2009) Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the b73 maize genome. PLOS Genetics 5(11): e1000732.

Benjamini, Y and Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics 29(4): 1165-1188.

Bolger, A M, Lohse M and Usadel B (2014) Trimmomatic: A flexible trimmer for illumina sequence data. Bioinformatics 30(15): 2114-2120.

Boratyn, G M, Camacho C, Cooper P S, Coulouris G, Fong A, Ma N, Madden T L, Matten W T, McGinnis S D, Merezhuk Y, Raytselis Y, Sayers E W, Tao T, Ye J and Zaretskaya I (2013) Blast: A more efficient report with usability improvements. Nucleic Acids Research 41(W1): W29-W33.

Cao, J, Schneeberger K, Ossowski S, Gunther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, Wang X, Ott F, Muller J, Alonso-Blanco C, Borgwardt K, Schmid K J and Weigel D (2011) Whole-genome sequencing of multiple arabidopsis thaliana populations. Nat Genet 43(10): 956-963.

Carvalho, M R, Herrera F A, Jaramillo C A, Wing S L and Callejas R (2011) Paleocene malvaceae from northern south america and their biogeographical implications. American Journal of Botany 98(8): 1337-1355.

Chia, J-M, Song C, Bradbury P J, Costich D, De Leon N, Doebley J, Elshire R J, Gaut B, Geller L and Glaubitz J C (2012) Maize hapmap2 identifies extant variation from a genome in flux. Nature genetics 44(7): 803-807.

Cronn, R C, Small R L, Haselkorn T and Wendel J F (2002) Rapid diversification of the cotton genus (gossypium: Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. American Journal of Botany 89(4): 707-725.

De La Torre, A R, Li Z, Van de Peer Y and Ingvarsson P K (2017) Contrasting rates of molecular evolution and patterns of selection among gymnosperms and flowering plants. Molecular Biology and Evolution 34(6): 1363-1377.

Dejoode, D R and Wendel J F (1992) Genetic diversity and origin of the hawaiian-islands cotton, gossypium-tomentosum. American Journal of Botany 79(11): 1311-1319.

DePristo, M A, Banks E, Poplin R, Garimella K V, Maguire J R, Hartl C, Philippakis A A, del Angel G, Rivas M A, Hanna M, McKenna A, Fennell T J, Kernytsky A M, Sivachenko A Y, Cibulskis K, Gabriel S B, Altshuler D and Daly M J (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43(5): 491-498.

Eddy, S R (2004) Where did the blosum62 alignment score matrix come from? Nat Biotech 22(8): 1035-1036.

Emms, D M and Kelly S (2015) Orthofinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biology 16(1): 157.

Ferreira de Carvalho, J, de Jager V, van Gurp T P, Wagemaker N C A M and Verhoeven K J F (2016) Recent and dynamic transposable elements contribute to genomic divergence under asexuality. BMC Genomics 17(1): 884.

Flagel, L E, Wendel J F and Udall J A (2012) Duplicate gene evolution, homoeologous recombination, and transcriptome characterization in allopolyploid cotton. BMC genomics 13(1): 302.

Flinders, A F, Ito G and Garcia M O (2010) Gravity anomalies of the northern hawaiian islands: Implications on the shield evolutions of kauai and niihau. Journal of Geophysical Research: Solid Earth 115(B8): n/a-n/a.

Fryxell, P A (1968) A redefinition of the tribe gossypieae. Botanical Gazette 129(4): 296-308.

Fryxell, P A (1979). The natural history of the cotton tribe (malvaceae, tribe gossypieae). College Station, Texas A&M University Press.

Gore, M A, Chia J-M, Elshire R J, Sun Q, Ersoz E S, Hurwitz B L, Peiffer J A, McMullen M D, Grills G S and Ross-Ibarra J (2009) A first-generation haplotype map of maize. Science 326(5956): 1115-1117.

Grabherr, M G, Haas B J, Yassour M, Levin J Z, Thompson D A, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren B W, Nusbaum C, Lindblad-Toh K, Friedman N and Regev A (2011) Trinity: Reconstructing a full-length transcriptome without a genome from rna-seq data. Nature biotechnology 29(7): 644-652.

Grandbastien, M A (2004) [stress activation and genomic impact of plant retrotransposons]. J Soc Biol 198(4): 425-432.

Gregory, T R (2003) Is small indel bias a determinant of genome size? Trends in Genetics 19(9): 485-488.

Gregory, T R and Witt J D S (2008) Population size and genome size in fishes: A closer look. Genome 51(4): 309-313.

Grover, C E, Hawkins J S and Wendel J F (2008a) Phylogenetic insights into the pace and pattern of plant genome size evolution. Genome Dyn 4: 57-68.

Grover, C E, Kim H, Wing R A, Paterson A H and Wendel J F (2004) Incongruent patterns of local and global genome size evolution in cotton. Genome Res 14(8): 1474-1482.

Grover, C E, Kim H, Wing R A, Paterson A H and Wendel J F (2007) Microcolinearity and genome evolution in the adha region of diploid and polyploid cotton (gossypium). Plant J 50(6): 995-1006.

Grover, C E, Yu Y, Wing R A, Paterson A H and Wendel J F (2008b) A phylogenetic analysis of indel dynamics in the cotton genus. Molecular Biology and Evolution 25(7): 1415-1428.

Gurevich, A, Saveliev V, Vyahhi N and Tesler G (2013) Quast: Quality assessment tool for genome assemblies. Bioinformatics 29(8): 1072-1075.

Hawkins, J S, Kim H, Nason J D, Wing R A and Wendel J F (2006) Differential lineage-specific amplification of transposable elements is responsible for genome size variation in gossypium. Genome Res 16(10): 1252-1261.

Hendrix, B and Stewart J M (2005) Estimation of the nuclear DNA content of gossypium species. Annals of Botany 95(5): 789-797.

Hirsch, C N, Foerster J M, Johnson J M, Sekhon R S, Muttoni G, Vaillancourt B, Peñagaricano F, Lindquist E, Pedraza M A, Barry K, de Leon N, Kaeppler S M and Buell C R (2014) Insights into the maize pan-genome and pan-transcriptome. The Plant Cell 26(1): 121-135.

Hjelmen, C E and Johnston J S (2017) The mode and tempo of genome size evolution in the subgenus sophophora. PLOS ONE 12(3): e0173505.

Holt, C and Yandell M (2011) Maker2: An annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12(1): 491.

Hutchinson, J (1943) A note on gossypium brevilanatum hochr. Trop Agric 20(4).

Hutchinson, J and Ghose R (1937) The composition of the cotton crops of central india and rajputana. Ind J Agric Sci 7(1).

Hutchinson, J B (1947) Notes on the classification and distribution of genera related to gossypium. New Phytologist 46(1): 123-141.

Kahlke, T, Goesmann A, Hjerde E, Willassen N P and Haugen P (2012) Unique core genomes of the bacterial family vibrionaceae: Insights into niche adaptation and speciation. BMC Genomics 13(1): 179.

Kalendar, R, Tanskanen J, Immonen S, Nevo E and Schulman A H (2000) Genome evolution of wild barley (hordeum spontaneum) by bare-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proc Natl Acad Sci U S A 97(12): 6603-6607.

Kapusta, A, Suh A and Feschotte C (2017) Dynamics of genome size evolution in birds and mammals. Proceedings of the National Academy of Sciences 114(8): E1460-E1469.

Kelly, L J, Renny-Byfield S, Pellicer J, Macas J, Novák P, Neumann P, Lysak M A, Day P D, Berger M, Fay M F, Nichols R A, Leitch A R and Leitch I J (2015) Analysis of the giant genomes of fritillaria (liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size. New Phytologist 208(2): 596-607.

Koch, M A, Haubold B and Mitchell-Olds T (2000) Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in arabidopsis, arabis, and related genera (brassicaceae). Molecular Biology and Evolution 17(10): 1483-1498.

Korf, I (2004) Gene finding in novel genomes. BMC Bioinformatics 5: 59-59.

Krzywinski, M, Schein J, Birol İ, Connors J, Gascoyne R, Horsman D, Jones S J and Marra M A (2009) Circos: An information aesthetic for comparative genomics. Genome Research 19(9): 1639-1645.

Lee, S-I and Kim N-S (2014) Transposable elements and genome size variations in plants. Genomics & Informatics 12(3): 87-97.

Lefébure, T, Morvan C, Malard F, François C, Konecny-Dupré L, Guéguen L, Weiss-Gayet M, Seguin-Orlando A, Ermini L, Sarkissian C D, Charrier N P, Eme D, Mermillod-Blondin F, Duret L, Vieira C, Orlando L and Douady C J (2017) Less effective selection leads to larger genomes. Genome Research.

Li, D, Liu C-M, Luo R, Sadakane K and Lam T-W (2015) Megahit: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. Bioinformatics 31(10): 1674-1676.

Liu, B and Wendel J F (2003) Epigenetic phenomena and the evolution of plant allopolyploids. Mol Phylogenet Evol 29(3): 365-379.

Lomsadze, A, Ter-Hovhannisyan V, Chernoff Y O and Borodovsky M (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Research 33(20): 6494-6506.

Lynch, M (2011) Statistical inference on the mechanisms of genome evolution. PLoS Genet 7(6): e1001389.

Lynch, M, Bobay L M, Catania F, Gout J F and Rho M (2011) The repatterning of eukaryotic genomes by random genetic drift. Annu Rev Genomics Hum Genet 12: 347-366.

Lynch, M and Conery J S (2003) The origins of genome complexity. Science 302(5649): 1401-1404.

McKenna, A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M and DePristo M A (2010) The genome analysis toolkit: A mapreduce framework for analyzing next-generation DNA sequencing data. Genome Research 20(9): 1297-1303.

Medini, D, Donati C, Tettelin H, Masignani V and Rappuoli R (2005) The microbial pan-genome. Current Opinion in Genetics & Development 15(6): 589-594.

Mohlhenrich, E R and Mueller R L (2016) Genetic drift and mutational hazard in the evolution of salamander genomic gigantism. Evolution 70(12): 2865-2878.

Morden, C W and Yorkston M (2017) Speciation and biogeography in the hawaiian endemic genus *kokia* (malvaceae: Gossypieae). Pacific Science in press.

Morgante, M, De Paoli E and Radovic S (2007) Transposable elements and the plant pan-genomes. Current Opinion in Plant Biology 10(2): 149-155.

Morton, B R, Gaut B S and Clegg M T (1996) Evolution of alcohol dehydrogenase genes in the palm and grass families. Proceedings of the National Academy of Sciences 93(21): 11735-11739.

Novák, P, Neumann P and Macas J (2010) Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics 11(1): 378.

Novák, P, Neumann P, Pech J, Steinhaisl J and Macas J (2013) Repeatexplorer: A galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics 29(6): 792-793.

Parisod, C, Alix K, Just J, Petit M, Sarilar V, Mhiri C, Ainouche M, Chalhoub B and Grandbastien M A (2010) Impact of transposable elements on the organization and function of allopolyploid genomes. New Phytol 186(1): 37-45.

Paterson, A H, Wendel J F, Gundlach H, Guo H, Jenkins J, Jin D, Llewellyn D, Showmaker K C, Shu S, Udall J, Yoo M-j, Byers R, Chen W, Doron-Faigenboim A, Duke M V, Gong L, Grimwood J, Grover C, Grupp K, Hu G, Lee T-h, Li J, Lin L, Liu T, Marler B S, Page J T, Roberts A W, Romanel E, Sanders W S, Szadkowski E, Tan X, Tang H, Xu C, Wang J, Wang Z, Zhang D, Zhang L, Ashrafi H, Bedon F, Bowers J E, Brubaker C L, Chee P W, Das S, Gingle A R, Haigler C H, Harker D, Hoffmann L V, Hovav R, Jones D C, Lemke C, Mansoor S, Rahman M u, Rainville L N, Rambani A, Reddy U K, Rong J-k, Saranga Y, Scheffler B E, Scheffler J A, Stelly D M, Triplett B A, Van Deynze A, Vaslin M F S, Waghmare V N, Walford S A, Wright R J, Zaki E A, Zhang T, Dennis E S, Mayer K F X, Peterson D G, Rokhsar D S, Wang X and Schmutz J (2012) Repeated polyploidization of gossypium genomes and the evolution of spinnable cotton fibres. Nature 492(7429): 423-427.

Paulino, D, Warren R L, Vandervalk B P, Raymond A, Jackman S D and Birol I (2015) Sealer: A scalable gap-closing application for finishing draft genomes. BMC Bioinformatics 16(1): 230.

Petrov, D A (2002) Mutational equilibrium model of genome size evolution. Theoretical Population Biology 61(4): 531-544.

Renny-Byfield, S, Page J T, Udall J A, Sanders W S, Peterson D G, Arick I I M A, Grover C E and Wendel J F (2016) Independent domestication of two old world cotton species. Genome Biology and Evolution 8(6): 1940-1947.

Richardson, J E, Whitlock B A, Meerow A W and Madriñán S (2015) The age of chocolate: A diversification history of theobroma and malvaceae. Frontiers in Ecology and Evolution 3(120).

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, Marçais G, Pop M and Yorke J A (2012) Gage: A critical evaluation of genome assemblies and assembly algorithms. Genome Research 22(3): 557-567.

Schnable, P S, Ware D, Fulton R S, Stein J C, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves T A, Minx P, Reily A D, Courtney L, Kruchowski S S, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock S M, Belter E, Du F, Kim K, Abbott R M, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson S M, Gillam B, Chen W, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado B, Mead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He R, Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin J, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania A, Ren L, Wei S, Kumari S, Faga B, Levy M J, McMahan L, Van Buren P, Vaughn M W, Ying K, Yeh C-T, Emrich S J, Jia Y, Kalyanaraman A, Hsia A-P, Barbazuk W B, Baucom R S, Brutnell T P, Carpita N C, Chaparro C, Chia J-M, Deragon J-M, Estill J C, Fu Y, Jeddeloh J A, Han Y, Lee H, Li P, Lisch D R, Liu S, Liu Z, Nagel D H, McCann M C, SanMiguel P, Myers A M, Nettleton D, Nguyen J, Penning B W, Ponnala L, Schneider K L, Schwartz D C, Sharma A, Soderlund C, Springer N M, Sun Q, Wang H, Waterman M, Westerman R, Wolfgruber T K, Yang L, Yu Y, Zhang L, Zhou S, Zhu Q, Bennetzen J L, Dawe R K, Jiang J, Jiang N, Presting G G, Wessler S R, Aluru S, Martienssen R A, Clifton S W, McCombie W R, Wing R A and Wilson R K (2009) The b73 maize genome: Complexity, diversity, and dynamics. Science 326(5956): 1112-1115.

Schwarz, G (1978) Estimating the dimension of a model. 461-464.

Seelanan, T, Schnabel A and Wendel J F (1997) Congruence and consensus in the cotton tribe (malvaceae). Systematic Botany 22(2): 259-290.

Senchina, D S, Alvarez I, Cronn R C, Liu B, Rong J, Noyes R D, Paterson A H, Wing R A, Wilkins T A and Wendel J F (2003) Rate variation among nuclear genes and the age of polyploidy in gossypium. Molecular Biology and Evolution 20(4): 633-643.

Service, U F a W. (2012). "Recovery plan for kokia cookei." 2017, from https://[www.fws.gov/pacificislands/flora/kokia.html](http://www.fws.gov/pacificislands/flora/kokia.html).

Sherwood, A R and Morden C W (2014) Genetic diversity of the endangered endemic hawaiian genus kokia (malvaceae). Pacific Science 68(4): 537-546.

Simão, F A, Waterhouse R M, Ioannidis P, Kriventseva E V and Zdobnov E M (2015) Busco: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31(19): 3210-3212.

Simpson, J T, Wong K, Jackman S D, Schein J E, Jones S J M and Birol İ (2009) Abyss: A parallel assembler for short read sequence data. Genome Research 19(6): 1117-1123.

Smit, A, Hubley R and Green P. (2013-2015). "Repeatmasker open-4.0." from <http://www.repeatmasker.org>.

Springer, N M, Ying K, Fu Y, Ji T, Yeh C-T, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H, Iniguez A L, Barbazuk W B, Jeddeloh J A, Nettleton D and Schnable P S (2009) Maize inbreds exhibit high levels of copy number variation (cnv) and presence/absence variation (pav) in genome content. PLOS Genetics 5(11): e1000734.

Stanke, M, Diekhans M, Baertsch R and Haussler D (2008) Using native and syntenically mapped cdna alignments to improve de novo gene finding. Bioinformatics 24(5): 637-644.

Stephens, S G (1958) Salt water tolerance of seeds of gossypium species as a possible factor in seed dispersal. American Naturalist 92(863): 83-92.

Stephens, S G (1966) The potentiality for long range oceanic dispersal of cotton seeds. The American Naturalist 100(912): 199-210.

Swanson-Wagner, R A, Eichten S R, Kumari S, Tiffin P, Stein J C, Ware D and Springer N M (2010) Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. Genome Research 20(12): 1689-1699.

Team, R C. (2017). "R: A language and environment for statistical computing." from https://[www.R-project.org/](http://www.R-project.org/).

Tettelin, H, Masignani V, Cieslewicz M J, Donati C, Medini D, Ward N L, Angiuoli S V, Crabtree J, Jones A L, Durkin A S, DeBoy R T, Davidsen T M, Mora M, Scarselli M, Margarit y Ros I, Peterson J D, Hauser C R, Sundaram J P, Nelson W C, Madupu R, Brinkac L M, Dodson R J, Rosovitz M J, Sullivan S A, Daugherty S C, Haft D H, Selengut J, Gwinn M L, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor K J B, Smith S, Utterback T R, White O, Rubens C E, Grandi G, Madoff L C, Kasper D L, Telford J L, Wessels M R, Rappuoli R and Fraser C M (2005) Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: Implications for the microbial “pan-genome”. Proceedings of the National Academy of Sciences of the United States of America 102(39): 13950-13955.

The UniProt Consortium (2017) Uniprot: The universal protein knowledgebase. Nucleic Acids Research 45(D1): D158-D169.

Tian, Z, Rizzon C, Du J, Zhu L, Bennetzen J L, Jackson S A, Gaut B S and Ma J (2009) Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? Genome Research 19(12): 2221-2230.

Van der Auwera, G A, Carneiro M O, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella K V, Altshuler D, Gabriel S and DePristo M A (2002). From fastq data to high-confidence variant calls: The genome analysis toolkit best practices pipeline.In: Current protocols in bioinformatics, John Wiley & Sons, Inc.

Vitte, C and Bennetzen J L (2006) Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proceedings of the National Academy of Sciences 103(47): 17638-17643.

Walker, B J, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo C A, Zeng Q, Wortman J, Young S K and Earl A M (2014) Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLOS ONE 9(11): e112963.

Wendel, J F (1989) New world tetraploid cottons contain old world cytoplasm. Proc Natl Acad Sci U S A 86(11): 4132-4136.

Wendel, J F and Albert V A (1992) Phylogenetics of the cotton genus (gossypium): Character-state weighted parsimony analysis of chloroplast-DNA restriction site data and its systematic and biogeographic implications. Systematic Botany 17(1): 115-143.

Wendel, J F and Cronn R C (2003). Polyploidy and the evolutionary history of cotton.In: Advances in agronomy, Academic Press. Volume 78**:** 139-186.

Wendel, J F, Cronn R C, Spencer Johnston J and James Price H (2002) Feast and famine in plant genomes. Genetica 115(1): 37-47.

Wendel, J F and Grover C E (2015) Taxonomy and evolution of the cotton genus, gossypium. Cotton(agronmonogr57): 25-44.

Wendel, J F and Percival A E (1990) Molecular divergence in the galapagos islands—baja california species pair,gossypium klotzschianum andg. Davidsonii (malvaceae). Plant Systematics and Evolution 171(1): 99-115.

Wendel, J F and Percy R G (1990) Allozyme diversity and introgression in the galapagos islands endemic gossypium darwinii and its relationship to continental g. Barbadense. Biochemical Systematics and Ecology 18(7): 517-528.

Whitney, K D, Boussau B, Baack E J and Garland T (2011) Drift and genome complexity revisited. Plos Genetics 7(6).

Whitney, K D and Garland T (2010) Did genetic drift drive increases in genome complexity? Plos Genetics 6(8).

Wickham, H (2016). Ggplot2: Elegant graphics for data analysis, Springer.

Wu, T D and Watanabe C K (2005) Gmap: A genomic mapping and alignment program for mrna and est sequences. Bioinformatics 21(9): 1859-1875.

Yang, Z (2007) Paml 4: Phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution 24(8): 1586-1591.

Yi, S and Streelman J T (2005) Genome size is negatively correlated with effective population size in ray-finned fish. Trends in Genetics 21(12): 643-646.