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

Authors and affiliations:

Corrinne E Grover[1*], Mark A Arick II[2], Justin L Conover[1], Adam Thrash[2], Guanjing Hu[1], William S Sanders[2,3,4], Chuan-Yu Hsu[2], Rubab Zahra Naqvi[5], Muhammad Farooq[5], Xiaochong Li[6], Lei Gong[6], Joann Mudge[7], Thiruvarangan Ramaraj[7], Joshua A Udall[8], Daniel G Peterson[2],and Jonathan F Wendel[1*]

[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, Faisalabad, Punjab, Pakistan

[6] Key Laboratory of Molecular Epigenetics of the Ministry of Education (MOE), Northeast Normal University, Changchun 130024, 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

* Corresponding authors

Data deposition: Genome assemblies and RNA-seq will be made available and listed by their BioProjects after processing through NCBI.

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 worldwide. 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 and use *de novo* genome sequences to address questions regarding the divergence of these two genera from each other and from their sister-group, *Gossypium*. In addition, we explore processes underlying the genome downsizing that characterizes *Kokia* and *Gossypioides* relative to other genera in the tribe. Using 13,000 gene orthologs and synonymous substitution rates, we show that the two disjuncts last shared a common ancestor about 5 MYA, or half as long ago as their divergence from *Gossypium*. 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. In both genera, there is a genome-wide bias toward deletions over insertions, and the number of gene losses exceeds the number of gains by about two- to four-fold. The genomic analyses presented here elucidate genomic consequences of the demographic and biogeographic history of these closest relatives of *Gossypium*, and enhance their value as phylogenetic outgroups.

Key words:

Gossypium; molecular evolution; genome evolution; genetic divergence; long-distance dispersal

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 occurred 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, i.e., *G. hirsutum* and *G. barbadense*). Outside of *Gossypium*, multiple long-distance dispersals have occurred during the evolution of the tribe (Dejoode and Wendel 1992; Fryxell 1979; Seelanan, et al. 1997; Stephens 1966, 1958; Wendel 1989; Wendel and Albert 1992; Wendel and Cronn 2003; Wendel and Percival 1990; Wendel and Percy 1990). 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 (Cronn, et al. 2002; Seelanan, et al. 1997).

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; Morden and Yorkston 2017; Sherwood and Morden 2014). 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 (Sherwood and Morden 2014; US Fish and Wildlife Service 2012). Due to the significance of *Kokia* to Hawaiian forests, diversity in the genus has been evaluated for the purposes of conservation (Morden and Yorkston 2017; Sherwood and Morden 2014). 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 (Morden and Yorkston 2017; Sherwood and Morden 2014).

The native region of *Gossypioides*, the sister genus to *Kokia*, is located over 17,500 kilometers distant in East Africa and Madagascar (Figure 1). 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* have a haploid chromosome base of 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 1943; Hutchinson and Ghose 1937). 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 (Fryxell 1968; Hutchinson 1947; 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; Lim and Marshall 2017), 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 the pre-release reference-quality sequence of *Gossypioides kirkii*. 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.

Material and 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 libraries were sequenced on a single lane of Illumina HiSeq2000 and two MiSeq flowcells (both at the Institute for Genomics, Biocomputing & Biotechnology, 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.

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. Each assembly was further scaffolded with ABySS using the MEGAHIT-derived transcripts. The assembly with the highest E-size (Salzberg, et al. 2012) was retained for improvement and analysis. ABySS Sealer v2.0.1 (Paulino, et al. 2015) was used to fill gaps in the retained 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. The *K. drynarioides* genome is available under NCBI BioProject PRJNA400144

Genome annotation of *K. drynarioides* and *G. kirkii*

RNA was extracted from three biological replicates of both *K. drynarioides* and *G. kirkii* for the purpose of annotation. Three centimeter (length) seedling leaves were collected and RNA as extracted using the Concert Plant RNA Reagent (Invitrogen) according to the manufacturer's instructions. Illumina libraries were generated 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. RNA-seq reads are available under BioProject PRJNA400144.

MAKER (v2.31.6) (Holt and Yandell 2011) annotation of the genome was then 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) (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 used to train Augustus and 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.

The pre-release genome sequence of *G. kirkii* (PRJNA400144) was similarly annotated (for assembly notes, please see https://github.com/Wendellab/KokiaKirkii). As with *K. drynarioides*, RNA was (1) extracted from three biological replicates of 3cm (length) *G. kirkii* seedling leaves using the Concert Plant RNA Reagent (Invitrogen) according to the manufacturer's instructions, (2) prepared for sequencing via TruSeq RNA Sample Preparation Kit (Illumina); and (3) sequenced as paired-end, 150 nt on the Illumina HiSeqX Ten at BerryGenomics (Beijing). MEGAHIT commit:02102e1 (Li, et al. 2015) was used to assemble the RNA data into transcripts, which were used in the same MAKER (Holt and Yandell 2011) iterations as used for *K. drynarioides* (see above). RNA-seq reads are available from NCBI PRJNA400144.

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.

Orthologous groups containing only a single representative from all species were retained (12,281 orthogroups out of 21,415 total, discarding approximately 17,000 genes from both *G. kirkii* and *K. drynarioides*). These retained groups were subsequently discarded if one or more representatives in that group contained ambiguous nucleotide bases (indicating poor sequence coverage; 50 groups total). 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 (Wang, et al. 2012). Distributions of all pairwise *dN*, *dS*, and *dN/dS* values were evaluated, and basic statistics (mean, median, and standard deviation) were calculated in R (R Core Team 2017).

Estimating Divergence Times

Absolute rates of synonymous substitutions have been estimated for eight angiosperm families (De La Torre, et al. 2017), fortuitously including the Malvaceae where the rate of substitution between *Theobroma* and *Gossypium* is 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) 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 in the sister genus, represented here by *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 independently completed by (1) delly2 (Rausch, et al. 2012) and (2) 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. Parameters for delly2 can be found at https://github.com/Wendellab/KokiaKirkii/tree/master/analysis/CNV_verification_delly2. Missing gene detection was completed by using genes annotated in *G. raimondii* that were not present in either *K. drynarioides* or *G. kirkii*, i.e., inferred losses, as gmap queries against the unmasked *K. drynarioides* or *G. kirkii*. Results were visualized with Circos (Krzywinski, et al. 2009).

Repeat clustering and annotation

All forward reads from the DNA libraries 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 (Hendrix and Stewart 2005; Wendel, et al. 2002) 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, 2007; Grover, et al. 2008b; Hawkins, et al. 2006; 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 (Supplementary Figure 1; 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 (R Core 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 (Altschul, et al. 1990; Boratyn, et al. 2013) 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 ($Y = a + bX + cX^2$), 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 (Supplementary Figure 2). 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 computed in R (R Core Team 2017), including the assumption of a 1:1 ratio 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 chi^2 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 DNA sequencing reads to the *G. raimondii* genome and using GATK (v 3.6) (DePristo, et al. 2011; McKenna, et al. 2010; Van der Auwera, et al. 2002) to align and characterize indels. GATK indel calls were pruned to remove (1) positions with missing data in either *G. kirkii* or *K. drynarioides* or (2) heterozygous sites. The resulting table was imported into R (R Core 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 (Supplementary Table 1; 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 (Supplementary Table 2). Annotation of the *K. drynarioides* genome (Supplementary Table 3) 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 pre-release *G. kirkii* genome (34X coverage PacBio; PRJNA400144) in the same manner as the *K. drynarioides* genome using two iterations of MAKER and the *G. kirkii* leaf RNA-seq generated here. 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 other two 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, Supplementary Table 4). Rates of molecular evolution among these three lineages were estimated for each singleton group (Supplementary Table 5), 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; Supplementary Table 6), whose median dS values were approximately equivalent (Figure 2). 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 (Figure 2; Supplementary Table 6).

Divergence Time within Malvaceae

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. 1997). More recently, divergence times for the Malvaceae, which includes cotton and chocolate, have been reported based on a single gene each from the chloroplast and the nuclear genomes (Richardson, et al. 2015), which suggests that chocolate (*Theobroma*) and *Gossypium* diverged *circa* 60-70 mya. Using a more extensive data set, absolute rates of synonymous substitutions have been estimated for eight angiosperm families (De La Torre, et al. 2017), fortuitously including the Malvaceae; this analysis estimates the rate of substitution between *Theobroma* and *Gossypium* as 4.56E–09/year.

Using this rate of substitution between *Theobroma* and *Gossypium*, whose fossil-based divergence time is approximately 60 mya, we revisit divergence times among *Kokia, Gossypioides, and Gossypium* using two orders of magnitude more genes (n = 13,643 single copy orthologs) extracted from the genome sequences for these taxa. The median of the resultant *dS* distribution (Supplementary Figure 3) was 0.4332, which predicts a synonymous substitution rate (r) of 3.61×10^{-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.

Gene 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. We inferred 731 genes gained and 2,957 lost in *G. kirkii* (distributed among 259 and 2,730 orthologous groups, respectively; Table 1). 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 1); 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. Thus, in both genera, the number of losses is about fourfold higher than the number of gains. 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 1). 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 (gains) or 0 (losses) in either *G. kirkii* or *K. drynarioides*). For the 211 gain events in *G. kirkii* and 394 in *K. drynarioides* meeting this criteria, few genes (1 - 8 %) were recovered from the remaining genome sequences (see methods), and in most cases, the predicted protein sequence was non-viable (Supplementary Table 4). 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, leading to an overall validation rate of 53.9%. 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 (64.8%

validation). Verification of deletions via delly2 led to a similar, but slightly lower overall validation rate for *G. kirkii* (40.8% of the original 2,957 inferred), but nearly complete validation of the original *K. drynarioides* deletion estimate (88.8%). The average number of deletions verified by the two independent methods suggests that the number of losses in both species exceeds the number of gains by about twofold since divergence (5.3 MYA) with a similar number of losses in both *G. kirkii* and *K. drynarioides*. Clearly this remarkable result will warrant further research, including using improved genomes and syntenic analyses to uncover the true extent and identity of gene deletions.

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

Because *K. drynarioides* and *G. kirkii* have relatively compact genomes (both 590 Mb), 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 the retained clusters represent a majority of the data set, i.e., most of the input data was represented in the analyzed clusters (Supplementary Figure 1).

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 (χ^2 $p > 0.95$). 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* (880 Mb) shows a highly distinct repetitive profile ($p < 0.05$) compared to either species from subgenus *Gossypium* (i.e., *G. herbaceum* and *G. arboreum*; 1667 Mb and 1689 Mb, respectively). 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 χ^2 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*, 94.5% of which are LTR-transposable elements (i.e., 61.8% *gypsy*, 10.9% *copia*, and 20% unspecified). Greater abundance was more frequently observed 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 (7,413 reads versus 7,252, 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 on 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” (Supplementary Table 7) and were annotated in approximately the same proportions as the

overall TE type classification. 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 3; (Baucom, et al. 2009; Hawkins, et al. 2006; Lee and Kim 2014; Paterson, et al. 2009; Schnable, et al. 2009; Tian, et al. 2009)). 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 2). Unsurprisingly, the small genomes of *K. drynarioides* and *G. kirkii* (590 Mb) had lower absolute quantities of most repeat types than the included diploid *Gossypium* genomes (i.e., *Gossypium raimondii*, 880 Mb; *G. arboreum*, 1689 Mb, and *G. herbaceum*, 1667 Mb) *except* for the non-LTR retrotransposon category. *K. drynarioides* and *G. kirkii* have comparable or slightly greater amounts of non-LTR retrotransposons as these three cotton species, despite the latter having 2-3x larger genomes (Figure 3). 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* (although present in different absolute amounts). 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 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 diploid cotton representatives as outgroup species (i.e., *Gossypium raimondii*, *G. arboreum*, and *G. herbaceum*). Patterns of both amplification and deletion were inferred (Figure 4), sometimes within the same cluster. For example, both *K. drynarioides* and *G. kirkii* have experienced reductions in copy number for repeat cluster 5 (*gypsy*),

albeit to different extents (Figure 4A). Likewise, the repeat represented by cluster 129 (*gypsy*) has experienced copy number growth in both *K. drynarioides* and *G. kirkii*, with the element attaining much higher copy numbers in *G. kirkii* (Figure 4B). A large subset of the repeat clusters (20 out of 22) showed gain in one of the two lineages coupled with concomitant loss in the other, creating differentially abundant clusters. Notably, no identifiable pattern of gain/loss or age bias was identified with respect to TE type. 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.

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 5). 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 6).

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, et al. 2002). 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 (Cronn, et al. 2002; Seelanan, et al. 1997). These initial estimates were from the pre-genomics era, and hence were based on relatively few nuclear and plastid genes. Here we present a updated estimate for the synonymous substitution rate (3.91×10^{-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 similar 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 (Cronn, et al. 2002; Seelanan, et al. 1997; Senchina, et al. 2003), which is remarkable observation given the fact that earlier estimates were based on two orders of magnitude fewer genes, although the caveat remains that these were derived from single representatives.

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 earliest emergence estimate for the present Hawaiian Islands. Because a signature trait of the Gossypieae is multiple trans-oceanic dispersals, these divergence data may suggest multiple trans-oceanic voyages between intermediate locale in the evolutionary history of *Kokia* (and any now-extinct members of its clade) before its arrival and diversification in the Hawaiian Islands concomitant with local extinction at any 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. Therefore, a credible alternative is that the antecedent of modern *Kokia* may have made a great leap circa 5.3 million years ago to the Hawaiian archipelago

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 (Fryxell 1979; Stephens 1958; 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; (Hirsch, et al. 2014; Medini, et al. 2005)). Research in plants (Cao, et al. 2011; Chia, et al. 2012; Hirsch, et al. 2014; Morgante, et al. 2007; Springer, et al. 2009; Swanson-Wagner, et al. 2010) suggests that many plant species exhibit evidence of a pan-genome whose "dispensable" component may contribute to diversity and adaptation (Kahlke, et al. 2012; Medini, et al. 2005; Tettelin, et al. 2005). Here, using a divergence time 5.3 million years, we estimate that gene deletions between *Kokia* and *Gossypioides* have occurred at about 245-300 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 outweigh insertions and identifiable sequence was not recovered from either *Kokia* or *Gossypioides*, we infer these missing sequences represent shared deletions that occurred in the ~5-6 MY between the divergence of *Gossypium* from proto-*Kokia*/*Gossypioides* and the divergence of the latter two genera from each other. While possible that incomplete assemblies in *K. drynarioides* and *G. kirkii* contributed to the observed rate of

gene loss, we note that the gene space for both is well-assembled (by BUSCO score) and it is improbable that both assemblies missed the same 6,000 gene models. The rate of gene deletion inferred for the shared lineage is much higher than in either individual lineage, 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 nearly equivalent. Notably, none of these lineages has a history of unshared paleopolyploidy (Paterson, et al. 2012), although all three share an ancient polyploidy event whose redundancy could be differentially fractionated.

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 (Lefébure, et al. 2017; Lynch 2011; Lynch, et al. 2011; Lynch and Conery 2003; Yi and Streelman 2005). 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 (Arnqvist, et al. 2015; Gregory and Witt 2008; Lefébure, et al. 2017; Mohlhenrich and Mueller 2016; Whitney, et al. 2011; Whitney and Garland 2010; Yi and Streelman 2005). 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 and stands in contrast to abundant literature that reports genome size differences among closely related species (Hawkins, et al. 2006; Kelly, et al. 2015; Macas, et al. 2015; Novak, et al. 2014; Piegu, et al. 2006; Tetreault and Ungerer 2016; Vu, et al. 2015) and even within species (Biemont 2008; Diez, et al. 2013; Duchoslav, et al. 2013; Smarda, et al. 2008). Furthermore, given the stresses associated with dispersal and colonization of new environments, it may be even more surprising that the genome size has remained the same; however, although biotic and abiotic stresses have been associated with TE release from suppression, this activation can be genotype dependent and repression under stress is also commonly observed (reviewed in Horvath, et al. 2017). Finally, research on plant invasiveness supports an association between small genome size and invasive potential, possibly due to the traits associated with invasive success (e.g., seedling growth rate, water/nutrient use, etc.; reviewed in Suda, et al., 2015). While these species are not invasive *per se*, colonization (such as that of the Hawaiian Islands and East Africa/Madagascar) may favor similar characteristics.

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 (Grandbastien 2004; Kalendar, et al. 2000; Liu and Wendel 2003; Parisod, et al. 2010). Ancestral state reconstructions of TE amounts (Figure 4) 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 and maybe indicative of nucleotypic or other phenotypic restrictions associated with genome size in these species.

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 (Gregory 2003; Hjelmen and Johnston 2017; Kapusta, et al. 2017; Petrov 2002; Vitte and Bennetzen 2006). 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. 2008a). 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.

The abundance of indels in these two small genomes may be a consequence of their presumably small effective population sizes during the course of evolution. An extension of the drift-barrier hypothesis to indel formation suggests that selection for DNA fidelity and repair is less efficient in small populations, leading to a higher rate of retained indel mutation events (Sung, et al. 2016). The predicted decline in DNA fidelity associated with decreasing population sizes has been broadly demonstrated across the tree of life (Sung, et al. 2016) and may explain the relative abundance of indels in these two lineages, including the increased propensity for indel formation in *K. drynarioides* which was likely subject to severe population size restrictions as it colonized the Hawaiian Islands. Differences in double-strand break repair have been invoked for genome size reduction in across the tree of life, including plant species (Hu, et al. 2011; Kirik, et al. 2000; Orel and Puchta 2003; Puchta 2005; Vu, et al. 2015).

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 similar 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/). Rubab Zahra Naqvi and Muhammad Farooq were funded by the "Pakistan-U.S. Cotton Productivity Enhancement Program" of ICARDA funded by the United States Department of Agriculture (USDA) Agricultural Research Service (ARS), under agreement No. 58-6402-0-178F. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the USDA or ICARDA. Partial funding was provided through USDA ARS agreement 58-6402-1-644 to Dan Peterson, the National Science Foundation (award #1145014to JAU and JFW), and Cotton Incorporated (to DGP).

Author contributions

C.E.G. and J.F.W conceived the project. C.E.G, M.A.A., W.S.S., R.Z.N., M.F., D.G.P, and J.F.W designed the experiments. G.H., C.H., X.L., L.G., J.M., T.R., and J.A.U. collected and prepared the data. C.E.G, M.A.A, J.L.C, and A.T. conducted the data analyses. C.E.G., M.A.A, J.L.C., D.G.P., and J.F.W wrote the manuscript. All authors revised the manuscript.

Figure 1: Modern geographic ranges of the genus *Kokia* in Hawaii and *Gossypioides* in East Africa/Madagascar, and their estimated divergence time (MYA = million years ago).

Figure 2: Distribution of substitution rates between pairwise comparisons of *K. drynarioides*, *G. kirkii*, and *Gossypium raimondii*. The line graph depicts the frequency distribution between *G. kirkii* and *K. drynarioides* (red); *G. kirkii* and *G. raimondii* (green); or *K. drynarioides* and *G. raimondii* (blue) calculated for 12,281 genes. Inset into the frequency graph are box plots of both the values (including the median) of both synonymous substitutions (red) and non-synonymous substitutions (black).

Figure 3: The (average) aggregate number of kilobases represented by each transposable element category for each species (genome sizes included next to species names). Transposable elements were broadly categorized into categories and their representation per species summarized. Multiple representatives were available for each *Gossypium* species (Renny-Byfield, et al. 2016), allowing an estimate of standard error for those species.

Figure 4: Example ancestral state reconstructions for gain/loss of sequence in the 22 clusters with the lowest p-value ($p<0.001$) during the evolution of *Kokia*/*Gossypioides*/*Gossypium*. The 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. (A) Exemplar cluster 5, *gypsy*; (B) Exemplar cluster 129, *gypsy*; (C) Exemplar cluster 110, *gypsy*; (D) Exemplar cluster 177, LTR-retrotransposon (unspecified type); and (E) Exemplar cluster 96, *gypsy*. Notably, no bias in TE type was detected for either gain/loss or age in *Kokia* and *Gossypioides*.

Figure 5: The frequency of indels present between *K. drynarioides* (green) and *G. kirkii* (blue), parsed as insertions (top) and deletions (bottom).

Figure 6: Genomic distribution of copy number variations and indels in *K. drynarioides* (Left) and *G. kirkii* (Right). **Ring 1**: gene gains (dark) and losses (light). **Ring 2**: insertions. **Ring 3**: deletions. **Ring 4:** mutual gene losses in *Kokia* and *Gossypioides*, relative to *Gossypium*.

Supplementary Figure 1: 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.

Supplementary Figure 2: 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 ($Y = a + bX + cX^2$) 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.

Supplemental Figure 3: 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 SF, Gish W, Miller W, Myers EW, Lipman DJ 1990. Basic local alignment search tool. Journal of Molecular Biology 215: 403-410. doi: http://dx.doi.org/10.1016/S0022-2836(05)80360-2

Arnqvist G, et al. 2015. Genome size correlates with reproductive fitness in seed beetles. Proceedings of the Royal Society B: Biological Sciences 282. doi: 10.1098/rspb.2015.1421

Bao W, Kojima KK, Kohany O 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA 6: 11. doi: 10.1186/s13100-015-0041-9

Bates DM. 1990. Malvaceae. In: Wagner W, Herbst D, Sohmer S, editors. Manual of the flowering plants of Hawai ‘i. Revised edition. Honolulu: University of Hawai'i and Bishop Museum Press. p. 868-902.

Baucom RS, et al. 2009. Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. Plos Genetics 5: e1000732. doi: 10.1371/journal.pgen.1000732

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

Biemont C 2008. Genome size evolution: within-species variation in genome size. Heredity 101: 297-298.

Bolger AM, Lohse M, Usadel B 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114-2120. doi: 10.1093/bioinformatics/btu170

Boratyn GM, et al. 2013. BLAST: a more efficient report with usability improvements. Nucleic Acids Research 41: W29-W33. doi: 10.1093/nar/gkt282

Cao J, et al. 2011. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nature genetics 43: 956-963. doi: http://www.nature.com/ng/journal/v43/n10/abs/ng.911.html#supplementary-information

Carvalho MR, Herrera FA, Jaramillo CA, Wing SL, Callejas R 2011. Paleocene Malvaceae from northern South America and their biogeographical implications. American Journal of Botany 98: 1337-1355. doi: 10.3732/ajb.1000539

Chia J-M, et al. 2012. Maize HapMap2 identifies extant variation from a genome in flux. Nature genetics 44: 803-807.

Cronn RC, Small RL, Haselkorn T, Wendel JF 2002. Rapid diversification of the cotton genus (*Gossypium*: Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. American Journal of Botany 89: 707-725. doi: 10.3732/ajb.89.4.707

De La Torre AR, Li Z, Van de Peer Y, Ingvarsson PK 2017. Contrasting rates of molecular evolution and patterns of selection among gymnosperms and flowering plants. Molecular Biology and Evolution 34: 1363-1377. doi: 10.1093/molbev/msx069

Dejoode DR, Wendel JF 1992. Genetic diversity and origin of the Hawaiian-Islands cotton, *Gossypium tomentosum*. American Journal of Botany 79: 1311-1319.

DePristo MA, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics 43: 491-498. doi: http://www.nature.com/ng/journal/v43/n5/abs/ng.806.html

Diez CM, Gaut BS, Meca E, Scheinvar E, Montes-Hernandez S, Eguiarte LE, Tenaillon MI 2013. Genome size variation in wild and cultivated maize along altitudinal gradients. New Phytoloist 199: 264-276. 10.1111/nph.12247

Duschoslav M, Safarova L, Jandova M 2013. Role of adaptive and non-adaptive mechanisms forming complex patterns of genome size variation in six cytotypes of polyploid *Allium oleraceum* (Amaryllidaceae) on a continental scale. Annals of Botany 111:419-431. https://doi.org/10.1093/aob/mcs297

Eddy SR 2004. Where did the BLOSUM62 alignment score matrix come from? Nature biotechnology 22: 1035-1036. doi: http://www.nature.com/nbt/journal/v22/n8/suppinfo/nbt0804-1035_S1.html

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

Ferreira de Carvalho J, de Jager V, van Gurp TP, Wagemaker NCAM, Verhoeven KJF 2016. Recent and dynamic transposable elements contribute to genomic divergence under asexuality. BMC Genomics 17: 884. doi: 10.1186/s12864-016-3234-9

Flagel LE, Wendel JF, Udall JA 2012. Duplicate gene evolution, homoeologous recombination, and transcriptome characterization in allopolyploid cotton. BMC Genomics 13: 302.

Flinders AF, Ito G, Garcia MO 2010. Gravity anomalies of the Northern Hawaiian Islands: Implications on the shield evolutions of Kauai and Niihau. Journal of Geophysical Research: Solid Earth 115: B08412. doi: 10.1029/2009JB006877

Fryxell PA. 1979. The natural history of the cotton tribe (Malvaceae, tribe Gossypieae). College Station: Texas A&M University Press.

Fryxell PA 1968. A redefinition of the tribe Gossypieae. Botanical Gazette 129: 296-308.

Grabherr MG, et al. 2011. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nature Biotechnology 29: 644-652. doi: 10.1038/nbt.1883

Grandbastien MA 2004. Journal de la Societe de Biologie. J Soc Biol 198: 425-432.

Gregory TR 2003. Is small indel bias a determinant of genome size? Trends in Genetics 19: 485-488. doi: http://dx.doi.org/10.1016/S0168-9525(03)00192-6

Gregory TR, Witt JDS 2008. Population size and genome size in fishes: a closer look. Genome 51: 309-313.

Grover CE, Hawkins JS, Wendel JF 2008a. Phylogenetic insights into the pace and pattern of plant genome size evolution. Genome Dyn 4: 57-68. doi: 10.1159/000126006

Grover CE, Kim H, Wing RA, Paterson AH, Wendel JF 2004. Incongruent patterns of local and global genome size evolution in cotton. Genome Research 14: 1474-1482. doi: 10.1101/gr.2673204

Grover CE, Kim H, Wing RA, Paterson AH, Wendel JF 2007. Microcolinearity and genome evolution in the AdhA region of diploid and polyploid cotton (*Gossypium*). Plant Journal 50: 995-1006. doi: 10.1111/j.1365-313X.2007.03102.x

Grover CE, Yu Y, Wing RA, Paterson AH, Wendel JF 2008b. A phylogenetic analysis of indel dynamics in the cotton genus. Molecular Biology and Evolution 25: 1415-1428. doi: 10.1093/molbev/msn085

Gurevich A, Saveliev V, Vyahhi N, Tesler G 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29: 1072-1075. doi: 10.1093/bioinformatics/btt086

Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF 2006. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. Genome Research 16: 1252-1261. doi: 10.1101/gr.5282906

Hendrix B, Stewart JM 2005. Estimation of the nuclear DNA content of *Gossypium* species. Annals of Botany 95: 789-797. doi: 10.1093/aob/mci078

Hirsch CN, et al. 2014. Insights into the maize pan-genome and pan-transcriptome. The Plant Cell 26: 121-135. doi: 10.1105/tpc.113.119982

Hjelmen CE, Johnston JS 2017. The mode and tempo of genome size evolution in the subgenus *Sophophora*. PLoS One 12: e0173505. doi: 10.1371/journal.pone.0173505

Holt C, Yandell M 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12: 491. doi: 10.1186/1471-2105-12-491

Horvath V, Merenciano M, Gonzalez J 2017. Revisiting the relationship between transposable elements and the eukaryotic stress response. Trends in Genetics 33: 832-841. http://dx.doi.org/10.1016/j.tig.2017.08.007

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng FJ, Clark RM, et al. 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nature Genetics 43: 476-481. doi:10.1038/ng.807

Hutchinson J 1943. A note on *Gossypium brevilanatum* Hochr. Tropical Agriculture 20.

Hutchinson J, Ghose R 1937. The composition of the cotton crops of Central India and Rajputana. Indian Journal of Agricultural Sciences 7.

Hutchinson JB 1947. Notes on the classification and distribution of genera related to *Gossypium*. New Phytologist 46: 123-141. doi: 10.1111/j.1469-8137.1947.tb05075.x

Kahlke T, Goesmann A, Hjerde E, Willassen NP, Haugen P 2012. Unique core genomes of the bacterial family vibrionaceae: insights into niche adaptation and speciation. BMC Genomics 13: 179. doi: 10.1186/1471-2164-13-179

Kalendar R, Tanskanen J, Immonen S, Nevo E, Schulman AH 2000. Genome evolution of wild barley (*Hordeum spontaneum*) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proceedings of the National Academy of Sciences USA 97: 6603-6607. doi: 10.1073/pnas.110587497

Kapusta A, Suh A, Feschotte C 2017. Dynamics of genome size evolution in birds and mammals. Proceedings of the National Academy of Sciences USA 114: E1460-E1469. doi: 10.1073/pnas.1616702114

Kelly LJ, et al. 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: 596-607. doi: 10.1111/nph.13471

Kirik A, Salomon S, Puchta H 2000. Species-specific double-strand break repair and genome evolution in plants. EMBO J. 19: 5562-5566.

Koch MA, Haubold B, 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: 1483-1498. doi: 10.1093/oxfordjournals.molbev.a026248

Korf I 2004. Gene finding in novel genomes. BMC Bioinformatics 5: 59-59. doi: 10.1186/1471-2105-5-59

Krzywinski M, et al. 2009. Circos: An information aesthetic for comparative genomics. Genome Research 19: 1639-1645. doi: 10.1101/gr.092759.109

Lee S-I, Kim N-S 2014. Transposable elements and genome size variations in plants. Genomics & Informatics 12: 87-97. doi: 10.5808/GI.2014.12.3.87

Lefébure T, et al. 2017. Less effective selection leads to larger genomes. Genome Research. doi: 10.1101/gr.212589.116

Li D, Liu C-M, Luo R, Sadakane K, Lam T-W 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31: 1674-1676. doi: 10.1093/bioinformatics/btv033

Lim JY, Marshall CR 2017. The true tempo of evolutionary radiation and decline revealed on the Hawaiian archipelago. Nature 543:710-713. doi:10.1038/nature21675

Liu B, Wendel JF 2003. Epigenetic phenomena and the evolution of plant allopolyploids. Mol Phylogenet Evol 29: 365-379.

Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Research 33: 6494-6506. doi: 10.1093/nar/gki937

Lynch M 2011. Statistical inference on the mechanisms of genome evolution. Plos Genetics 7: e1001389. doi: 10.1371/journal.pgen.1001389

Lynch M, Bobay LM, Catania F, Gout JF, Rho M 2011. The repatterning of eukaryotic genomes by random genetic drift. Annu Rev Genomics Hum Genet 12: 347-366. doi: 10.1146/annurev-genom-082410-101412

Lynch M, Conery JS 2003. The origins of genome complexity. Science 302: 1401-1404.

Macas J, Novak P, Pellicer J, Cizkova J, Koblizkova A, Neumannn P, Fukova I, Dolezel J, Kelly LJ, Leitch IJ. In Depth Characterization of Repetitive DNA in 23 Plant Genomes Reveals Sources of Genome Size Variation in the Legume Tribe *Fabeae*. PLOS ONE 10:e0143424. https://doi.org/10.1371/journal.pone.0143424

McKenna A, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20: 1297-1303. doi: 10.1101/gr.107524.110

Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R 2005. The microbial pan-genome. Current Opinion in Genetics & Development 15: 589-594. doi: https://doi.org/10.1016/j.gde.2005.09.006

Mohlhenrich ER, Mueller RL 2016. Genetic drift and mutational hazard in the evolution of salamander genomic gigantism. Evolution 70: 2865-2878. doi: 10.1111/evo.13084

Morden CW, Yorkston M 2017. Speciation and biogeography in the Hawaiian endemic genus *Kokia* (Malvaceae: Gossypieae). Pacific Science in press.

Morgante M, De Paoli E, Radovic S 2007. Transposable elements and the plant pan-genomes. Current Opinion in Plant Biology 10: 149-155. doi: http://dx.doi.org/10.1016/j.pbi.2007.02.001

Morton BR, Gaut BS, Clegg MT 1996. Evolution of alcohol dehydrogenase genes in the palm and grass families. Proceedings of the National Academy of Sciences USA 93: 11735-11739.

Novak P, Hribova E, Neumann P, Koblizkova A, Dolezel J, Macas J. Genome-wide analysis of repeat diversity across the family Musaceae. PLOS ONE 9:e98918. https://doi.org/10.1371/journal.pone.0098918

Novák P, Neumann P, Macas J 2010. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics 11: 378. doi: 10.1186/1471-2105-11-378

Novák P, Neumann P, Pech J, Steinhaisl J, Macas J 2013. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics 29: 792-793. doi: 10.1093/bioinformatics/btt054

Orel N, Puchta H 2003. Differences in the processing of DNA ends in *Arabidopsis thaliana* and tobacco: possible implications for genome evolution. Plant Molecular Biology 51:523-531.

Parisod C, et al. 2010. Impact of transposable elements on the organization and function of allopolyploid genomes. New Phytologist 186: 37-45. doi: 10.1111/j.1469-8137.2009.03096.x

Paterson AH, et al. 2009. The *Sorghum bicolor* genome and the diversification of grasses. Nature 457: 551-556. doi: http://www.nature.com/nature/journal/v457/n7229/suppinfo/nature07723_S1.html

Paterson AH, et al. 2012. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. Nature 492: 423-427. doi: http://www.nature.com/nature/journal/v492/n7429/abs/nature11798.html#supplementary-information

Paulino D, et al. 2015. Sealer: a scalable gap-closing application for finishing draft genomes. BMC Bioinformatics 16: 230. doi: 10.1186/s12859-015-0663-4

Petrov DA 2002. Mutational equilibrium model of genome size evolution. Theoretical Population Biology 61: 531-544. doi: http://dx.doi.org/10.1006/tpbi.2002.1605

Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, Collura K, Brar DS, Jackson S, Wing RA, Panaud O 2006. Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. Genome Research 16: 1262-1269. 10.1101/gr.5290206

Puchta H 2005. The repair of double-strand breaks in plants: mechanisms and consequences for genome evolution. Journal of Experimental Botany 56:1-14.

R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2017. Available from: https://www.R-project.org/

Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28:i333-i339. 10.1093/bioinformatics/bts378

Renny-Byfield S, et al. 2016. Independent domestication of two Old World cotton species. Genome Biol Evol 8: 1940-1947. doi: 10.1093/gbe/evw129

RepeatMasker Open-4.0 [Internet]. 2013-2015. Available from: http://www.repeatmasker.org

Richardson JE, Whitlock BA, Meerow AW, Madriñán S 2015. The age of chocolate: a diversification history of *Theobroma* and Malvaceae. Frontiers in Ecology and Evolution 3. doi: 10.3389/fevo.2015.00120

Salzberg SL, et al. 2012. GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Research 22: 557-567. doi: 10.1101/gr.131383.111

Schnable PS, et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. Science 326: 1112-1115. doi: 10.1126/science.1178534 https://doi.org/10.1016/j.tplants.2016.06.003

Schubert I, Vu GTH 2016. Genome stability and evolution: attempting a holistic view. Trends in Plant Science 21:749-757.

Schwarz G 1978. Estimating the dimension of a model. 461-464. doi: 10.1214/aos/1176344136

Seelanan T, Schnabel A, Wendel JF 1997. Congruence and consensus in the cotton tribe (Malvaceae). Systematic Botany 22: 259-290. doi: 10.2307/2419457

Senchina DS, et al. 2003. Rate variation among nuclear genes and the age of polyploidy in *Gossypium*. Molecular Biology and Evolution 20: 633-643. doi: 10.1093/molbev/msg065

Sherwood AR, Morden CW 2014. Genetic diversity of the endangered endemic Hawaiian genus *Kokia* (Malvaceae). Pacific Science 68: 537-546. doi: 10.2984/68.4.7

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210-3212. doi: 10.1093/bioinformatics/btv351

Simpson JT, et al. 2009. ABySS: A parallel assembler for short read sequence data. Genome Research 19: 1117-1123. doi: 10.1101/gr.089532.108

Smarda P, Bures P, Horova L, Rotreklova O 2008. Intrapopulations genome size dynamics in *Festuca pallens*. Annals of Botany 102:559-607. https://doi.org/10.1093/aob/mcn133

Springer NM, et al. 2009. Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. Plos Genetics 5: e1000734. doi: 10.1371/journal.pgen.1000734

Stanke M, Diekhans M, Baertsch R, Haussler D 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24: 637-644. doi: 10.1093/bioinformatics/btn013

Stephens SG 1966. The potentiality for long range oceanic dispersal of cotton seeds. The American Naturalist 100: 199-210.

Stephens SG 1958. Salt water tolerance of seeds of *Gossypium s*pecies as a possible factor in seed dispersal. The American Naturalist 92: 83-92.

Suda J, Meyerson LA, Leitch IJ, Pysek P 2014. The hidden side of plant invasions: the role of genome size. New Phytologist 205:994-1007. 10.1111/nph.13107

Sung W, Ackerman MS, Dillon MM, Platt TG, Fuqua C, Cooper VS, Lynch M 2016. Evolution of the insertion-deletion mutation rate across the tree of life. G3 Genes|Genomes|Genetics 6: 2583-2591. doi: 10.1534/g3.116.030890

Swanson-Wagner RA, et al. 2010. Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. Genome Research 20: 1689-1699. doi: 10.1101/gr.109165.110

Tetreault HM, Ungerer MC 2016. Long terminal repeat retrotransposon content in eight diploid sunflower species inferred from next-generation sequence data. G3 Genes|Genomes|Genetics 6:2299-2308. https://doi.org/10.1534/g3.116.029082

Tettelin H, et al. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome". Proceedings of the National Academy of Sciences USA 102: 13950-13955. doi: 10.1073/pnas.0506758102

The UniProt Consortium 2017. UniProt: the universal protein knowledgebase. Nucleic Acids Research 45: D158-D169. doi: 10.1093/nar/gkw1099

Tian Z, et al. 2009. Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? Genome Research 19: 2221-2230. doi: 10.1101/gr.083899.108

Recovery plan for *Kokia cookei* [Internet]. Portland, USA2012 [cited 2017. Available from: https://www.fws.gov/pacificislands/flora/kokia.html

Van der Auwera GA, et al. 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, Bennetzen JL 2006. Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proceedings of the National Academy of Sciences USA 103: 17638-17643. doi: 10.1073/pnas.0605618103

Vu GTH, Schmutzer T, Bull F, Cao HX, Fuchs J, Tran TD, Jovtchev G, Pistrick K, Stein Nils, Pecinka A, Neumann P, Novak P, Macas J, Dear PH, Blattner FR, Scholz U, Schubert I 2015. Comparative genome analysis reveals divergent genome size evolution in a carnivorous plant genus. The Plant Genome 8:1-14. doi:10.3835/plantgenome2015.04.0021.

Walker BJ, et al. 2014. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9: e112963. doi: 10.1371/journal.pone.0112963

Wang K, et al. 2012. The draft genome of a diploid cotton *Gossypium raimondii*. Nature genetics 44: 1098-1103. doi: http://www.nature.com/ng/journal/v44/n10/abs/ng.2371.html#supplementary-information

Wendel JF 1989. New World tetraploid cottons contain Old World cytoplasm. Proceedings of the National Academy of Sciences USA 86: 4132-4136.

Wendel JF, Albert VA 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: 115-143. doi: 10.2307/2419069

Wendel JF, Cronn RC. 2003. Polyploidy and the evolutionary history of cotton. In. Advances in Agronomy: Academic Press. p. 139-186.

Wendel JF, Cronn RC, Spencer Johnston J, James Price H 2002. Feast and famine in plant genomes. Genetica 115: 37-47. doi: 10.1023/a:1016020030189

Wendel JF, Grover CE 2015. Taxonomy and evolution of the cotton genus, *Gossypium*. Cotton: 25-44.

Wendel JF, Percival AE 1990. Molecular divergence in the Galapagos Islands—Baja California species pair, *Gossypium klotzschianum* and *G. davidsonii* (Malvaceae). Plant Systematics and Evolution 171: 99-115. doi: 10.1007/BF00940598

Wendel JF, Percy RG 1990. Allozyme diversity and introgression in the Galapagos Islands endemic Gossypium darwinii and its relationship to continental G. barbadense. Biochemical Systematics and Ecology 18: 517-528. doi: http://dx.doi.org/10.1016/0305-1978(90)90123-W

Whitney KD, Boussau B, Baack EJ, Garland T 2011. Drift and genome complexity revisited. Plos Genetics 7.

Whitney KD, Garland T 2010. Did genetic drift drive increases in genome complexity? Plos Genetics 6.

Wu TD, Watanabe CK 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21: 1859-1875. doi: 10.1093/bioinformatics/bti310

Yang Z 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution 24: 1586-1591. doi: 10.1093/molbev/msm088

Yi S, Streelman JT 2005. Genome size is negatively correlated with effective population size in ray-finned fish. Trends in Genetics 21: 643-646.

Figure 1

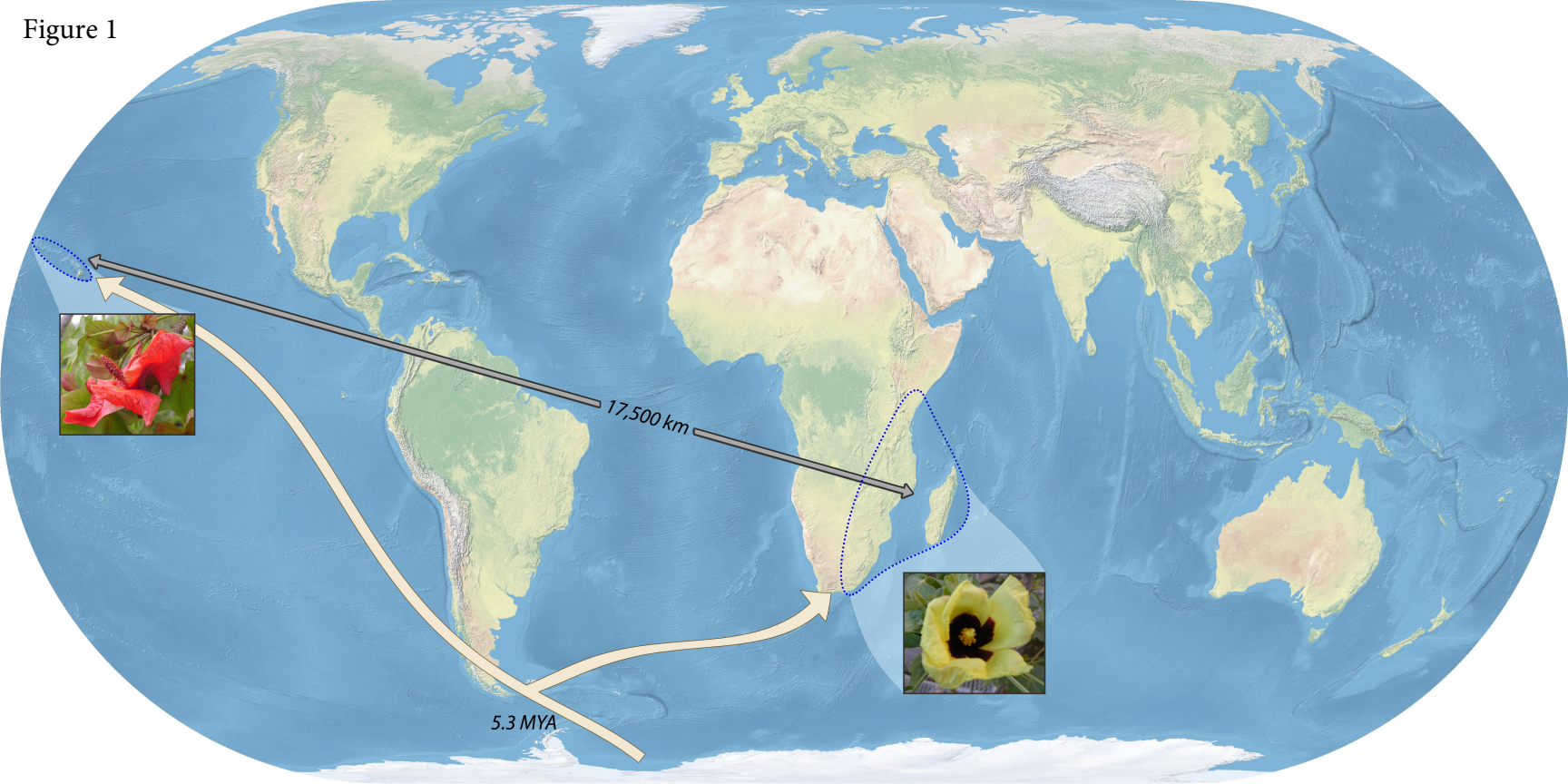


Figure 2

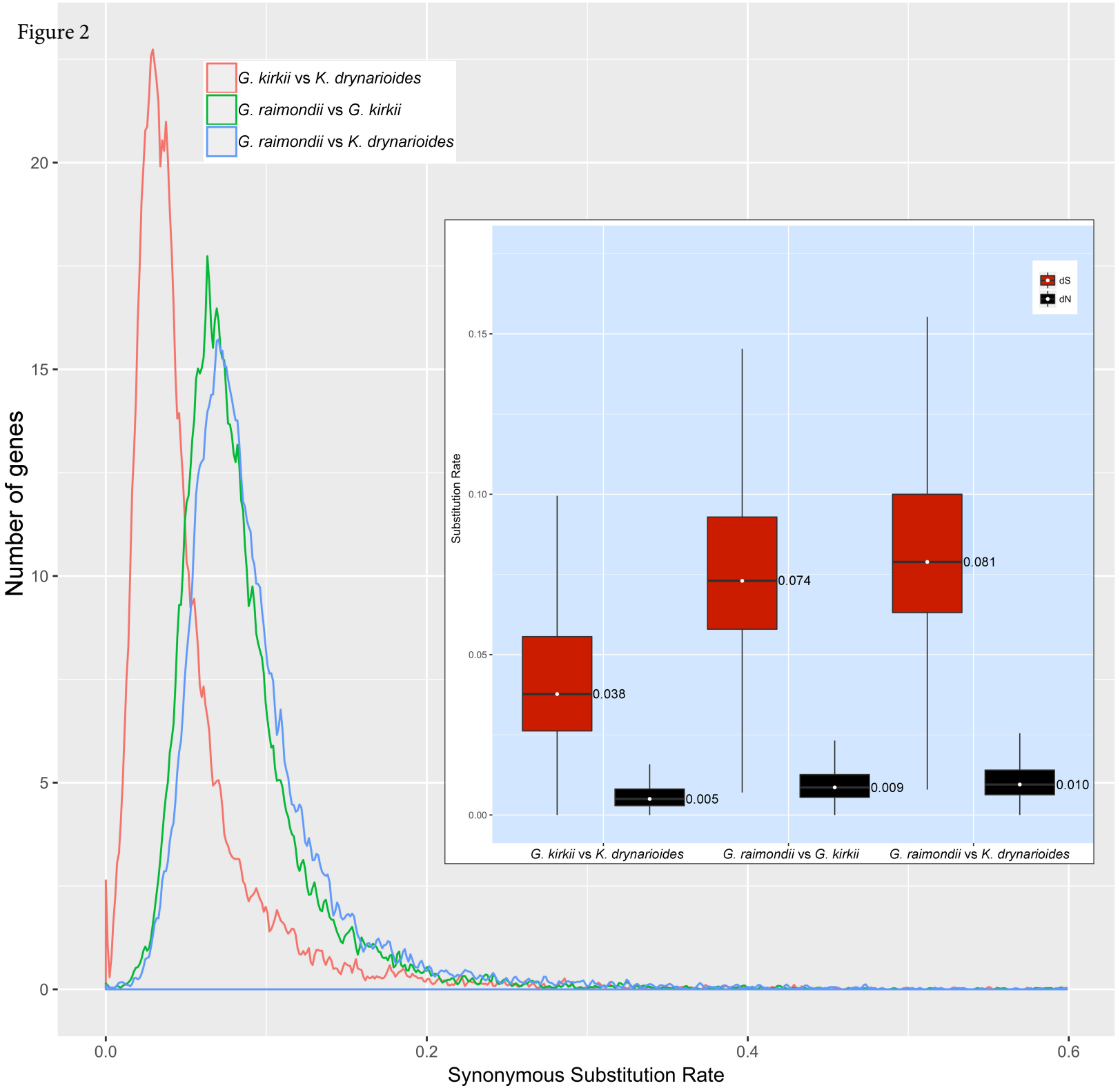
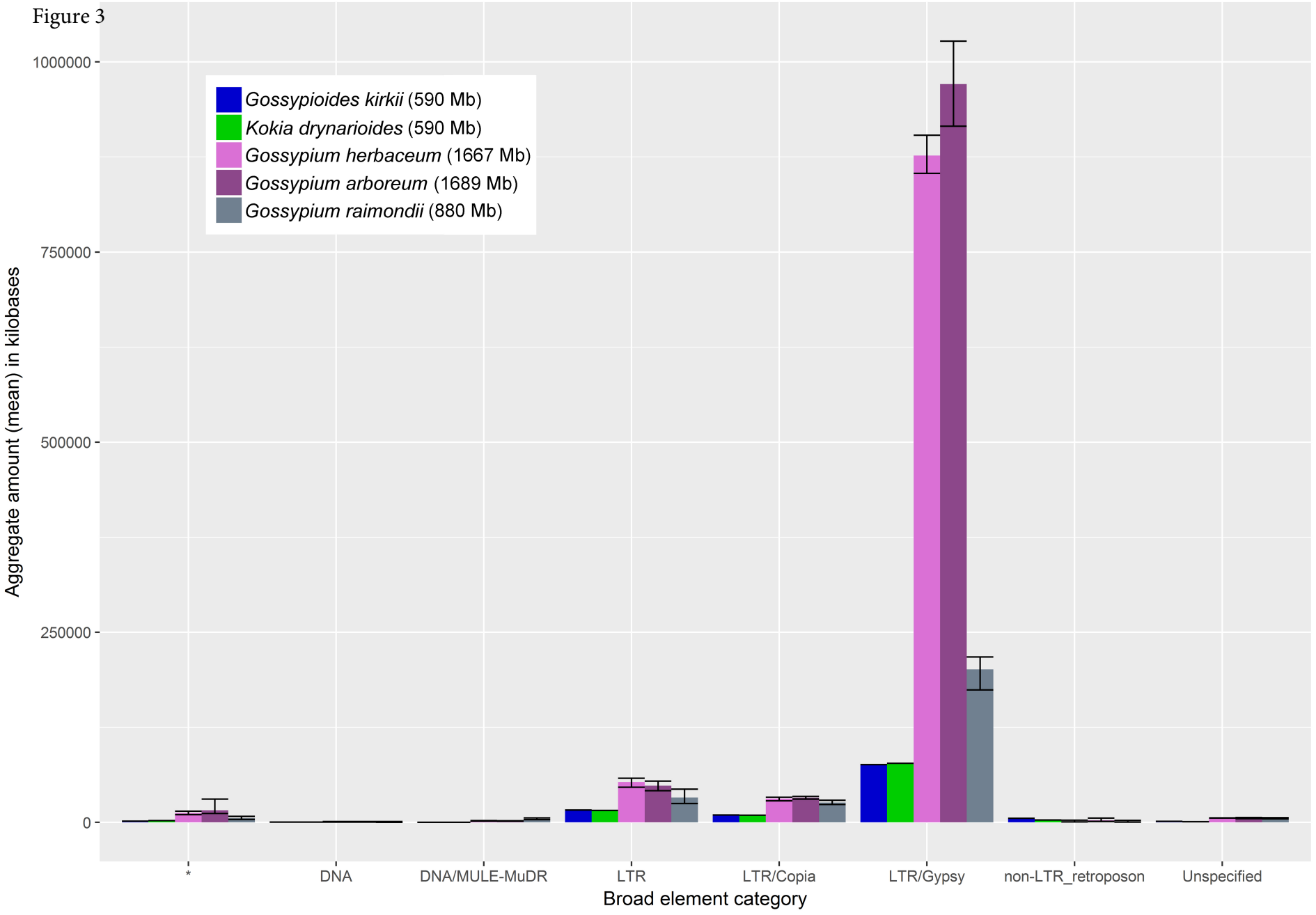
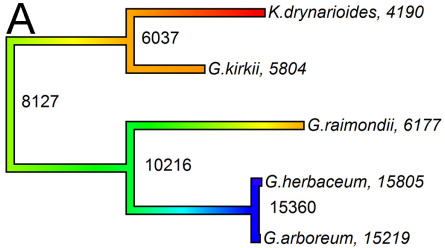
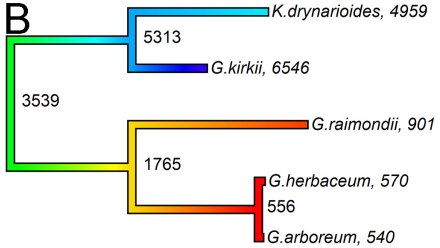
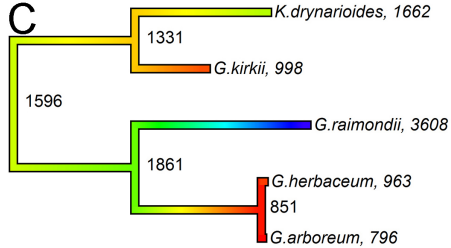

G. kirkii vs K. drynarioides
G. raimondii vs G. kirkii
G. raimondii vs K. drynarioides
Number of genes
Synonymous Substitution Rate
0
5
10
15
20
0.0
0.2
0.4
0.6
Substitution Rate
0.00
0.05
0.10
0.15
dS
dN
0.038
0.005
0.074
0.009
0.081
0.010
G. kirkii vs K. drynarioides
G. raimondii vs G. kirkii
G. raimondii vs K. drynarioides

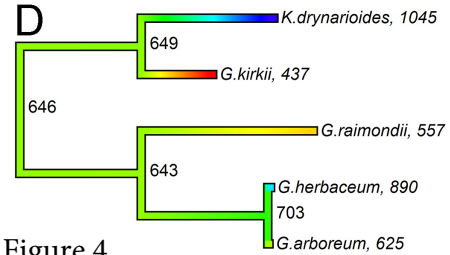
Figure 3

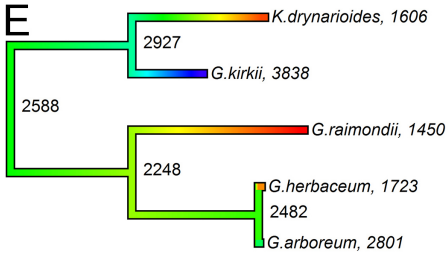

Gossypioides kirkii (590 Mb)
Kokia drynarioides (590 Mb)
Gossypium herbaceum (1667 Mb)
Gossypium arboreum (1689 Mb)
Gossypium raimondii (880 Mb)
1000000
750000
500000
250000
0
Aggregate amount (mean) in kilobases
*
DNA
DNA/MULE-MuDR
LTR
LTR/Copia
LTR/Gypsy
non-LTR_retroposon
Unspecified
Broad element category











Genome Sizes

K. drynarioides 590 Mb
G. kirkii 590 Mb
G. raimondii 880 Mb
G. herbaceum 1667 Mb
G. arboreum, 1689 Mb

Figure 4

Figure 5

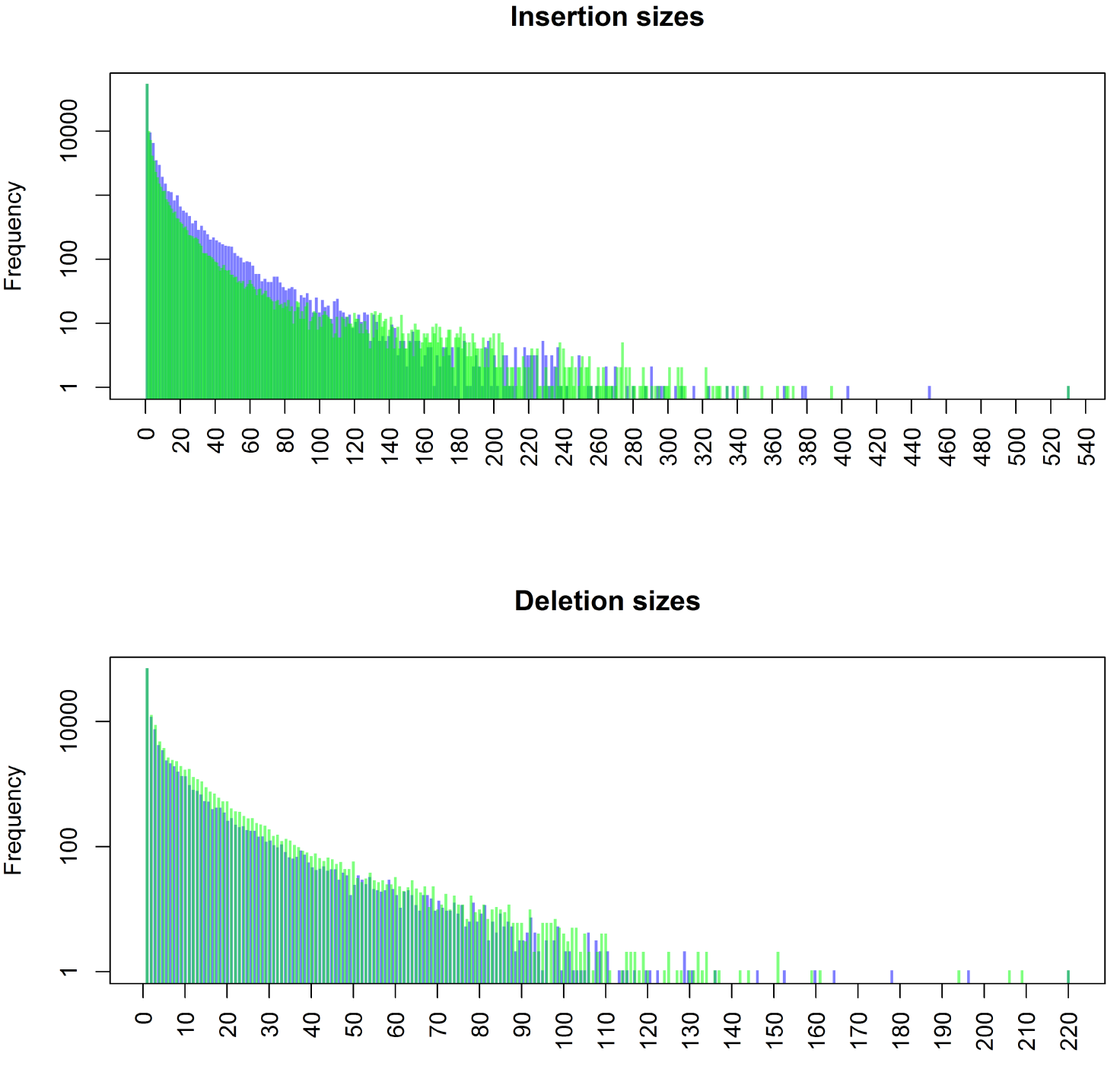
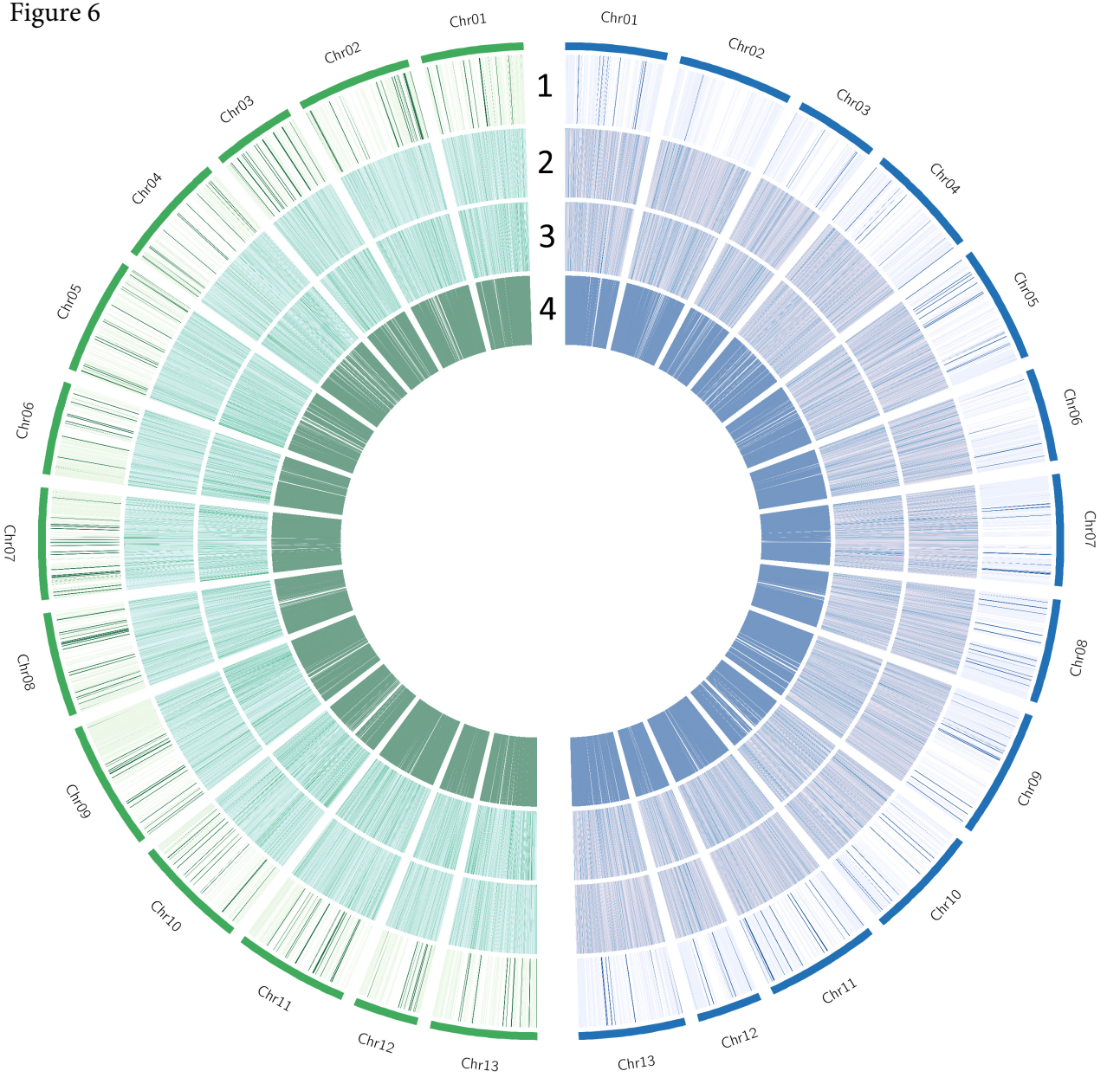

Insertion sizes
Frequency
1
10
100
10000
0
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
320
340
360
380
400
420
440
460
480
500
520
540
Deletion sizes
Frequency
1
100
10000
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220

Figure 6


Chr01
Chr02
Chr03
Chr04
Chr05
Chr06
Chr07
Chr08
Chr09
Chr10
Chr11
Chr12
Chr13
1
2
3
4

		Size of Gains/Losses in Orthogroup											
		1	2	3	4	5	6	7	8	14*	225*	Total Events	Total Genes
Gain	G. kirkii	222	23	5	3	2	2	2	0	1	1	259	731
	K. drynarioides	260	200	33	3	1	1	0	1	0	0	499	790
Loss	G. kirkii	2551	145	25	6	2	0	1	0	0	0	2730	2957
	K. drynarioides	1857	58	7	2	0	1	0	0	0	0	1925	2008

Table 1:

*These groups were not used in calculation of Total Events or Total Genes. We infer these genes are either falsely annotated Transposable Elements or and error in the clustering algorithm used.

Table 2: Aggregate amount of each broad TE category per genome (in kb)

Lineage	***K. drynarioides***	***G. kirkii***	***G. herbaceum***	***G. arboreum***	***G. raimondii***
Unspecified	2,413	1,520	12,299	15,842	5,267
DNA	219	190	823	610	447
DNA/MULE-MuDR	86	76	1,967	1,784	4,592
LTR	15,770	16,207	52,890	48,241	32,619
LTR/Copia	9,491	9,719	31,401	32,283	26,110
LTR/Gypsy	77,596	76,000	876,955	970,883	200,992