

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

Corrinne E Grover¹, Mark A Arick II¹, Justin Conover¹, Adam Thrasher¹, Guanqing Hu¹, William S Sanders¹, Rubab Naqvi¹, Muhammad Farooq¹, Joann Mudge¹, Thiru Ramaraj¹, Joshua A Udall¹, Daniel G Peterson¹, Jodi Scheffler¹, Brian Scheffler¹, Jonathan F Wendel¹

1 Affiliation Dept/Program/Center, Institution Name, City, State, Country

2 Affiliation Dept/Program/Center, Institution Name, City, State, Country

3 Affiliation Dept/Program/Center, Institution Name, City, State, Country

*** E-mail: Corresponding author@institute.edu**

Abstract

Author Summary

1 Introduction

2 One of the intriguing phenomena that characterizes the cotton tribe, Gossypieae, is the
3 prevalence of long-distance, trans-oceanic dispersals. The most famous of these occur
4 within the cotton genus itself (*Gossypium*); however, multiple events are found
5 throughout the tribe [1–10]. The sister genera *Kokia* and *Gossypoides* both represent a
6 minimum of one such oceanic dispersal followed by individual regional speciation. Based
7 on molecular divergence estimates derived from both chloroplast and nuclear genes,
8 these genera collectively diverged from the cotton genus during the Miocene
9 approximately 10-15 million years ago (mya; [10,11]), subsequently splitting into
10 individual genera and achieving widely dispersed, yet very localized ranges.

11 *Kokia* (Malvaceae) is a small Hawaiian endemic genus composed of four species that
12 were once widespread, major components of Hawaiian forests, yet are now all either
13 endangered, or recently extinct (*K. lanceolata* Lewton; [12,13]). Few individuals remain
14 of the two free-living extant species, *K. kauaiensis* (Rock) Degener & Duvel and *K.*
15 *drynarioides* (Seem.) Lewton, the latter of which is critically endangered and nearly
16 extinct in the wild, while the third endangered species, *K. cookei* Degener, exists only as
17 a maintained graft derived from a single individual ([13,14]). The native region of its
18 sister genus, *Gossypoides*, is located over 15,000 kilometers away in East Africa and
19 Madagascar. The two species that comprise the genus, *G. kirkii* M. Mast. and *G.*
20 *brevilanatum* Hoch. (East Africa and Madagascar, respectively), are themselves
21 reproductively isolated and, with *Kokia*, are cytologically distinct from the remainder of
22 the cotton tribe in that they appear to have experienced an aneuploid reduction in
23 chromosome number. Specifically, while most genera in the Gossypieae are based on
24 $n=13$, species in both *Kokia* and *Gossypoides* are $n=12$, likely representing a
25 chromosome loss or fusion event. The two species of *Gossypoides* also are
26 cytogenetically distinct, with an unusually long chromosome pair in *G.*
27 *brevilanatum* [15,16].

28 Despite the extensive research on the evolution of *Gossypium*, these sister genera
29 have been grossly understudied, except in serving as phylogenetic outgroups for cotton

30 phylogenetic and genomic research [10, 11]. Genomic resources in both genera are
 31 minimal, access to plant material is limited, and with the recent exception of a study by
 32 Sherwood and Morden (2014) on diversity among *Kokia* species, much of our knowledge
 33 regarding these genera is decades old [10, 17, 18].

34 The history of these genera, however, is intriguing. The current distribution of *Kokia*
 35 in the Hawaiian Islands and *Gossypioides* in East Africa-Madagascar necessitates at
 36 least one significant trans-oceanic traversal to a relatively young island chain that began
 37 to emerge only about 3.4 mya, an age approximately equivalent to the estimated
 38 divergence between *Kokia drynarioides* and *Gossypioides kirkii* [10] and slightly more
 39 recent than the basal most divergence in *Gossypium*. Diversity within *Gossypioides* is
 40 unknown, aside from acquisition of reproductive isolation between its sole two species;
 41 however, diversity in *Kokia* has been evaluated for the purposes of conservation [13]. A
 42 remarkable amount of diversity within and among species has been detected,
 43 particularly given the demographic history of *Kokia*, which includes the original genetic
 44 bottleneck of the founder, range expansion, and the subsequent bottleneck of habitat
 45 loss and the introduction of competitive and/or damaging alien species [13].

46 Direct comparisons of these genera are limited. Hutchinson (1943) notes that
 47 successful grafts can be made between *Kokia drynarioides* and *Gossypioides kirkii*, and
 48 their shared chromosomal reduction ($n=12$) is unique in the tribe. Estimates using a
 49 small number of nuclear genes suggest that genic distance between *K. drynarioides* and
 50 *G. kirkii* are similar to estimates between basally diverged species in *Gossypium*, i.e.,
 51 approximately 2% versus 3%, although a slight increase in replacement site
 52 substitutions is observed [11].

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

60 **Materials and Methods**

61 ***Kokia drynarioides* sequencing and genome assembly**

62 DNA was extracted from mature leaves using the Qiagen Plant DNeasy kit (Qiagen).
 63 Total genomic DNA was independently sheared via HOW into two average sizes, i.e.,
 64 350bp and 550bp, for Illumina library construction. A single, independent libraries was
 65 constructed from each fragment pool using the Illumina PCR-free library construction
 66 kit (Illumina). The 350 bp library was sequenced on a single lane of Illumina HiSeq2000
 67 and the larger, 550bp library was sequenced on two MiSeq flowcells (both at IGBB,
 68 Mississippi State University).

69 The data were trimmed and filtered with Trimmomatic v0.32 [19] with the following
 70 options: (1) sequence adapter removal, (2) removal of leading and/or trailing bases
 71 when the quality score (Q) <28, (3) removal of bases after average Q <28 (8 nt window)
 72 or single base quality <10, and (4) removal of reads <85 nt.

73 RNA was extracted MEGAHIT commit:02102e1 [20] was used to assemble the
 74 RNA data into transcripts.

75 The trimmed DNA data and RNA assembly were assembled via ABySS v2.0.1 [21],
 76 using every 5th kmer value from 65 through 200. The assembly with the highest
 77 E-size [22] was retained for improvement and analysis. Each retained assembly was
 78 further scaffolded with ABySS using the MEGAHIT-derived transcripts. ABySS Sealer

79 v2.0.1 [23] was used to fill gaps in the scaffolded assembly using every 10th kmer
80 starting at 100 and decreasing to 30. Pilon v1.22 [24] polished the resulting gap-filled
81 assembly using all trimmed DNA data. QUAST v4.5 [25] was used to generate the final
82 assembly statistics. (Let's get this all into github)

83 Genome annotation

84 MAKER (v2.31.6) [26] annotation of the genome was completed in two rounds, using
85 only contigs of <1 kb and training MAKER with Kokia-specific sequences. First, pass
86 de novo annotations were derived from Genemark (v4.3.3) [27] and retained for
87 MAKER training. At the same time, BUSCO (v2) [28] was used both to train Augustus
88 and create a Snap model^{Corrinne}. Finally, Trinity^{Corrinne} (v2.2.0) [29] was used to create
89 an RNASeq-assembly to pass to MAKER as EST evidence. The first pass of MAKER
90 was run using the combination of: (1) the output from Genemark, (2) the
91 BUSCO-generated Snap model, (3) the BUSCO-trained Augustus [30] model, (4) the
92 Trinity RNASeq-assembly as ESTs, and (5) the UniProt protein database.

Corrinne: WHAT'S A
SNAP MODEL

Corrinne: WHY
TRINITY VS
MEGAHIT

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

99 Identification of Orthologs

100 Amino acid sequences from *G. kirkii*, *G. raimondii* and *K. drynarioides* were clustered
101 using OrthoFinder v1.1.41 [31], which utilizes a Markov clustering algorithm of
102 normalized BLASTp scores to infer homology between proteins sequences of different
103 species. OrthoFinder is similar to OrthoMCL2 [32], but reduces the number of BLAST
104 results by filtering scores based on reciprocal best hits (RBHs) and corrects for gene
105 length biases and floor-limitation of e-values in BLAST scores prior to clustering. These
106 corrections have been shown to increase precision by improved clustering of singletons
107 (i.e., groups in which only one gene from each species is present) instead of entire gene
108 families into a given orthologous group. Default values were used for the inflation
109 parameter (1.5) in the Markov clustering, and the “-og” flag was used to prevent
110 downstream analyses after the groups were generated.

111 dN/dS Estimation and Timing of Divergence

112 Singletons inferred from OrthoFinder were separated into all 3 possible pairwise groups
113 (Gr + Gk, Gr + Kd, Kd + Gk). Amino acid sequences from each pairwise group were
114 then aligned using the pairwise2 python package and the BLOSUM62 substitution
115 matrix. The highest scoring alignments were then used as a guide to codon-align the
116 CDS sequences. The CODEML package in PAML [33] was used to calculate the dN, dS,
117 and dN/dS values. Singletons in which any pairwise comparison resulted in a dS value
118 greater than 0.03^{JustinCorrinne} was removed from the analysis and inferred to be a
119 cluster of non-orthologous proteins. Distributions of all pairwise dN, dS, and dN/dS
120 values were then plotted, and mean value and standard deviation is reported. Estimates
121 of total divergence time between each pairwise group was calculated using the equation
122 $T = dS / (2r)$ where r is the absolute rate of synonymous substitutions of *Adh* genes in
123 palms (2.6×10^{-9} substitutions X substitution site⁻¹ X year⁻¹) [11, 34] or members of
124 Brassicaceae (1.5×10^{-8} substitutions X synonymous site⁻¹ X year⁻¹) [35].

Justin: May need to
adjust after doing
said analysis

Corrinne: What was
our justification for
this again?

125 Copy Number Variation Estimation

126 A custom Python script (<https://github.com/Wendellab/KokiaKirkii>) was used to
127 calculate lineage-specific gene losses and duplications as inferred by OrthoFinder. A
128 gene loss was defined as an orthologous group in which 2 species had the same number
129 of genes present (n), but the third species contained $n-1$ genes. Likewise, a gene
130 duplication was identified by 2 species containing n genes, while the third contained
131 $n+1$. [JustinCorrinne](#)

Justin: Very rough estimate of gene loss and duplication; do we want more sophisticated method? Other parts to this section?

132 Repeat clustering and annotation

133 All reads from one of the paired-end files (i.e., R1) were filtered for quality and trimmed
134 to a standard 95nt using Trimmomatic version 0.33 [19] as per
135 (<https://github.com/Wendellab/KokiaKirkii>). Surviving reads were randomly
136 subsampled to represent a 1% genome size equivalent for each genome [36,37] and
137 combined as input into the RepeatExplorer pipeline [38,39], which is designed to cluster
138 reads based on similarity and identify putative repetitive sequences using low-coverage,
139 small read sequencing. Clusters containing a minimum of 0.01% of the total input
140 sequences (i.e., 201 reads from a total input of 2,013,469 reads) were annotated by the
141 RepeatExplorer implementation of RepeatMasker [40] using a custom library derived
142 from a combination of Repbase version WHATEVER [41] and previously annotated
143 cotton repeats [42–46]. A cutoff of 0.01% read representation is common; however, we
144 evaluated the suitability of this cut using a log of diminishing returns (FIGURE
145 WHATEVER; <https://github.com/Wendellab/KokiaKirkii>).

Corrinne: We probably should cross-check these to make sure things didn't get screwed up, e.g., a gene "loss" is actually where something got thrown in as a "duplication" or as a loner (true singleton with no match in other genomes)

146 Within the annotated clusters, the number of megabases (Mb) attributable to that
147 cluster (i.e., element type) for each genome/accession was calculated based on the 1%
148 genome representation of the sample and the standardized read length of 95 nt; total
149 repetitive amounts for each broad repetitive classification were summed from these
150 results. The genome occupation of each cluster (i.e., the calculated number of Mb) was
151 normalized by genome size for each accession, resulting in the percent of each genome
152 occupied by that element type, for use in multivariate visualization (i.e., Principle
153 Coordinate Analysis and Principal Component Analysis). All analyses were conducted
154 in R [47]; R versions and scripts are available at
155 (<https://github.com/Wendellab/KokiaKirkii>).

156 Repeat heterogeneity and relative age

157 Relative cluster age was approximated using the among-read divergence profile of each
158 cluster, as previously used for *Fritillaria* [48] and dandelion [49]. Briefly, an all-versus-all
159 BLASTn [50,51] was conducted on a cluster-by-cluster basis using the same BLAST
160 parameters implemented in RepeatExplorer. A histogram of pairwise percent identity
161 was generated for each cluster and the trend (i.e., biased toward high-identity, "young"
162 or lower-identity, "older" element reads) was described for each via regression models
163 using R. Specifically, two regression models were used to describe the data as either
164 linear ($Y = a + bX$) or quadratic ($Y = a + bX + cX^2$), and the model with the highest
165 confidence was determined via Bayesian Information Criterion [52]. The read similarity
166 profile for each cluster was automatically evaluated for each histogram to determine if
167 the reads trend toward highly similar "young" or more divergent "older" reads, as per
168 (Julie paper) with an additional category. These categories include (1) positive linear
169 regression; (2) absence of linear regression; (3) negative linear regression; (4) positive
170 quadratic vertical parabola, trend described by right-side of vertex; (4b) positive
171 quadratic vertical parabola, trend described by left-side of vertex; (5) negative quadratic
172 vertical parabola, trend described by right-side of vertex; and (6) negative quadratic

vertical parabola, trend described by left-side of vertex and vertex at 99% pairwise-identity (Figure WHATEVER). Categories which trend toward highly identical reads (i.e., 1, 4, and 6) were interpreted as having relatively young membership, whereas categories which trend toward lower identity (i.e., 2, 3, 4b, and 5) were interpreted as being composed of older elements. As with Ferreira de Carvalho (2016), this regression simply provides a relative characterization of cluster/element age and is not designed to detect statistically significant differences.

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

Comparison of abundance for the annotated clusters in *Kokia drynarioides* and *Gossypoides kirkii* were visualized via ggplot [53], 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 [54].

Results

Kokia genome assembly and annotation

Table 1. *Kokia* Genome Assembly Statistics. All statistics are based on contigs of size ≥ 500 bp, unless otherwise noted (e.g., "# contigs (≥ 0 bp)" and "Total length (≥ 0 bp)" include all contigs).

Assembly	<i>Kokia</i> Scaffolds	<i>Kokia</i> Contigs
# contigs (≥ 0 bp)	130430	-
# contigs (≥ 1000 bp)	15404	21494
# contigs (≥ 5000 bp)	7390	11998
# contigs (≥ 10000 bp)	5267	8994
# contigs (≥ 25000 bp)	3543	5501
# contigs (≥ 50000 bp)	2385	2984
Total length (≥ 0 bp)	537779651	-
Total length (≥ 1000 bp)	518114202	516998315
Total length (≥ 5000 bp)	499433075	494744281
Total length (≥ 10000 bp)	484390717	473215614
Total length (≥ 25000 bp)	456973742	415721086
Total length (≥ 50000 bp)	415141424	325473417
# contigs	19146	25827
Largest contig	2291099	974327
Total length	520833981	520152831
GC (%)	33.08	33.08
N50	176649	72430
N75	66795	31815
L50	756	1895
L75	1960	4594
# N's per 100 kbp	84.02	2.00

Table 2. BUSCO (Single-Copy Orthologs) Statistics

Type	Count
Complete BUSCOs	1377
Complete and single-copy BUSCOs	1213
Complete and duplicated BUSCOs	164
Fragmented BUSCOs	17
Missing BUSCOs	46
Total BUSCO groups searched	1440

Table 3. Kokia Annotation Statistics.

Feature	Total Predicted	W/ Evidence	W/ Strong Evidence
Genes	29231		
mRNAs	29231		
CDSs	171914		

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

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

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

Because *K. drynarioides* and *G. kirkii* have relatively compact genomes, multiple representatives of three cotton species previously used for repetitive analysis [55] were included in the clustering to aid in the identification of repeat-derived sequences. Just over two million reads derived from these five species (comprising 1% genome size equivalents each) were co-clustered using the RepeatExplorer pipeline, producing a total 74,001 clusters (n \geq 2 reads). Because the smallest clusters are neither informative nor reliable indicators of repetitiveness, we chose to annotate only those clusters composed of greater than 0.01% of the total reads input (=201 reads), resulting in 274 retained clusters. We evaluated the cumulative read sum as the cluster number increases (clusters are numbered from largest to smallest) to confirm that this represents a reasonable partitioning of the data set.

Despite identically sized genomes, *K. drynarioides* and *G. kirkii* show an approximately 1 Mb difference in clustered repeats, although this lacks statistical significance. Contingency table analysis of the repetitive profiles of each species, as well as the total amount of repetitive DNA calculated for each, suggest that

: cotton_cutoff.png

Corrinne: put the linear regression stuff in here?

these profiles are indistinguishable (at $p < 0.05$), despite the intergeneric comparison. Interspecies (intra-genus) repetitive profiles for those *Gossypium* species present in the analysis showed a different pattern, whereby the basally divergent *G. raimondii* compared to either A-genome species (i.e., *G. herbaceum* and *G. arboreum*) shows a highly distinct repetitive profile ($p < 0.05$), although, notably, the sister A-genome species are not distinct (see discussion).

To ascertain the extent of the differences between *K. drynarioides* and *G. kirkii*, we considered the possibility that while the overall repetitive profiles may not be significantly different, individual clusters may be. Toward this end, we conducted a chi² test of independence for each cluster and applied a Benjamini-Hochberg correction for multiple testing. At $p < 0.05$, XXX clusters (out of XXX) are differentially abundant in *K. drynarioides* versus *G. kirkii*, with the species displaying greater abundance occurring approximately the same number of times for both (XXX with greater abundance in *K. drynarioides* versus XXX in *G. kirkii*; Table Abundance). Because these differentially abundant clusters could represent differences in either proliferation or decay/removal, we gauged the relative age of each cluster based of the method of Ferreira de Carvalho (2016). This analysis attempts to characterize the age of each cluster^{Corrinne} based on the distinctiveness of the reads which comprise the cluster; that is, younger clusters will have reads that are highly similar, whereas older clusters will have reads that show a number of differences. While an imperfect measure, this characterization permits a generalized perspective on the repeats identified here. Overall, most of the repeats in *K. drynarioides* and *G. kirkii* displayed a pattern suggestive of older elements (202 versus 72 “young”); however, of the XXX differentially abundant clusters, XXX were categorized as “young” and XXX as “older” (Table Abundance), potentially reflecting SOMETHING ABOUT GAIN VERSUS LOSS.

Corrinne: should we redo this just for the Kok/Kirk reads? would the A-genome reads, minimally, be biasing some of these toward “youth”?

Most of the clusters were broadly annotated as belonging to the Ty3/gypsy superfamily, a result not surprising for a plant lineage (Figure Amounts). Overall, gypsy elements comprise XXX to XXX of the *K. drynarioides* and *G. kirkii* genomes, respectively, with uncategorized LTR-retrotransposons and Ty1/copia elements comprising the next most abundant repeats and comprising similar amounts in each genome. Unsurprisingly, the small genomes of *K. drynarioides* and *G. kirkii* had lower absolute abundance of most repeat types except the predicted non-LTR retrotransposons, in which these two species had comparable or slightly greater occupation as the cotton species, which possess 2-3x larger genomes. This difference is due to the sole retrotransposon clusters recovered, which was in the top 5 largest clusters for both *K. drynarioides* and *G. kirkii*. The high percent identity among reads for this cluster suggests it is relatively young, and it has likely experienced proliferation in both species. Furthermore, the cluster shows differential abundance between the two species, suggesting that either the proliferation began prior to species divergence and continued with varying success afterwards, or the two lineages experienced similar releases from repression for this element, although again to varying degrees. The other differentially abundant clusters were largely annotated as putative gypsy elements (RIGHT?) (XX %).

Discussion

Divergence and speciation are expected outcomes of long-distance insular dispersal, whose conceptual foundations are rooted in the observations of Darwin and other early evolutionary biologists. The tribe Gossypieae is characterized by such dispersals, ultimately achieving worldwide distribution on all tropical and subtropical-inclusive continents. Most Gossypieae genera, save for the eponymous *Gossypium* (cotton genus), have been grossly understudied except as each pertains to the evolution of cotton. Here we present first-pass genome assemblies for the outgroup congeners to *Gossypium*,

270 which together provide insight into the interesting biogeographic history of these genera
271 and their equivocality as outgroups in studying the evolution of the cotton genus.

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

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

276 Long-distance salt water dispersal common in gossypieae
277 Advance Agronomy

278 • *Lebronnecia* – marquesas (south pacific)

279 • *Thespecia thespesioides* – pan tropical

280 • *Hampia* – neotropical (americas)

281 • *Thespesia populnea* – pan tropical

282 • *Cephalohibiscus* – new guinea and solomon islands (Australia)

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

289 Supporting Information

290 S1 Video

291 **Bold the first sentence.** Maecenas convallis mauris sit amet sem ultrices gravida.
292 Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula.
293 Curabitur fringilla pulvinar lectus consectetur pellentesque.

294 S1 Text

295 **Lorem Ipsum.** Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget
296 sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur
297 fringilla pulvinar lectus consectetur pellentesque.

298 S1 Fig

299 **Lorem Ipsum.** Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget
300 sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur
301 fringilla pulvinar lectus consectetur pellentesque.

302 S1 Table

303 **Lorem Ipsum.** Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget
304 sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur
305 fringilla pulvinar lectus consectetur pellentesque.

306 Acknowledgments

307 Cras egestas velit mauris, eu mollis turpis pellentesque sit amet. Interdum et malesuada
308 fames ac ante ipsum primis in faucibus. Nam id pretium nisi. Sed ac quam id nisi
309 malesuada congue. Sed interdum aliquet augue, at pellentesque quam rhoncus vitae.

References

1. DeJoo DR, Wendel JF (1992) Genetic diversity and origin of the hawaiian-islands cotton, *gossypium-tomentosum*. American Journal of Botany 79: 1311-1319.
2. Fryxell PA (1979) The natural history of the cotton tribe (Malvaceae, tribe Gossypieae). College Station: Texas A&M University Press, 1st edition, xviii, 245 p. pp. 78021779 by Paul A. Fryxell. ill. ; 24 cm. Bibliography: p. [227]-232. Includes index. Cotton tribe.
3. Stephens SG (1958) Salt water tolerance of seeds of *gossypium* species as a possible factor in seed dispersal. American Naturalist 92: 83-92.
4. Stephens SG (1966) The potentiality for long range oceanic dispersal of cotton seeds. The American Naturalist 100: 199-210.
5. Wendel JF (1989) New world tetraploid cottons contain old world cytoplasm. Proc Natl Acad Sci U S A 86: 4132-6.
6. 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.
7. 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.
8. 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.
9. Wendel JF, Cronn RC (2003) Polyploidy and the evolutionary history of cotton, Academic Press, volume Volume 78. pp. 139-186. doi:10.1016/S0065-2113(02)78004-8.
10. Seelanan T, Schnabel A, Wendel JF (1997) Congruence and consensus in the cotton tribe (malvaceae). Systematic Botany 22: 259-290.
11. 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.
12. Bates DM (1990) Malvaceae, Honolulu: University of Hawai'i and Bishop Museum Press. pp. 868-902.
13. Sherwood AR, Morden CW (2014) Genetic diversity of the endangered endemic hawaiian genus *kokia* (malvaceae). Pacific Science 68: 537-546.
14. Service UF, Wildlife (2012). Recovery plan for *kokia cookei*.

15. Hutchinson J, Ghose R (1937) The composition of the cotton crops of central india and rajputana. *Ind J Agric Sci* 7.
16. Hutchinson J (1943) A note on gossypium brevilanatum hochr. *Trop Agric* 20.
17. Hutchinson JB (1947) Notes on the classification and distribution of genera related to gossypium. *New Phytologist* 46: 123-141.
18. Fryxell PA (1968) A redefinition of the tribe gossypieae. *Botanical Gazette* 129: 296-308.
19. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 30: 2114-2120.
20. Li D, Liu CM, Luo R, Sadakane K, Lam TW (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics* 31: 1674-1676.
21. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, et al. (2009) ABySS: A parallel assembler for short read sequence data. *Genome Research* 19: 1117-1123.
22. Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, et al. (2012) GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research* 22: 557-567.
23. Paulino D, Warren RL, Vandervalk BP, Raymond A, Jackman SD, et al. (2015) Sealer: a scalable gap-closing application for finishing draft genomes. *BMC Bioinformatics* 16.
24. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, et al. (2014) Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* 9: e112963.
25. Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29: 1072-1075.
26. Holt C, Yandell M (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 12: 491.
27. Lomsadze A (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Research* 33: 6494-6506.
28. 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.
29. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, et al. (2011) Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnology* 29: 644-652.
30. Stanke M, Waack S (2003) Gene prediction with a hidden markov model and a new intron submodel. *Bioinformatics* 19: ii215-ii225.
31. Emms DM, Kelly S (2015) Orthofinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* 16: 157.

32. Li L, Stoeckert CJ, Roos DS (2003) Orthomcl: Identification of ortholog groups for eukaryotic genomes. *Genome Research* 13: 2178-2189.
33. Yang Z (2007) Paml 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* 24: 1586-1591.
34. 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* 93: 11735-11739.
35. 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.
36. Hendrix B, Stewart JM (2005) Estimation of the nuclear dna content of gossypium species. *Annals of Botany* 95: 789-797.
37. Wendel JF, Cronn RC, Spencer Johnston J, James Price H (2002) Feast and famine in plant genomes. *Genetica* 115: 37-47.
38. 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.
39. Novák P, Neumann P, Macas J (2010) Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics* 11: 378.
40. Smit A, Hubley R, Green P (2013-2015). Repeatmasker open-4.0.
41. Bao W, Kojima KK, Kohany O (2015) Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* 6: 11.
42. Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, et al. (2012) Repeated polyploidization of gossypium genomes and the evolution of spinnable cotton fibres. *Nature* 492: 423-427.
43. Grover CE, Hawkins JS, Wendel JF (2008) Phylogenetic insights into the pace and pattern of plant genome size evolution. *Genome Dyn* 4: 57-68.
44. 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 J* 50: 995-1006.
45. Grover CE, Kim H, Wing RA, Paterson AH, Wendel JF (2004) Incongruent patterns of local and global genome size evolution in cotton. *Genome Res* 14: 1474-82.
46. 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 Res* 16: 1252-61.
47. Team RC (2017). R: A language and environment for statistical computing.
48. Kelly LJ, Renny-Byfield S, Pellicer J, Macas J, Novák P, 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.

49. 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.
50. Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, et al. (2013) Blast: a more efficient report with usability improvements. *Nucleic Acids Research* 41: W29-W33.
51. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* 215: 403-410.
52. Schwarz G (1978) Estimating the dimension of a model. *The Annals of Statistics* : 461-464.
53. Wickham H (2016) *ggplot2: elegant graphics for data analysis*. Springer.
54. Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* 29: 1165-1188.
55. Renny-Byfield S, Page JT, Udall JA, Sanders WS, Peterson DG, et al. (2016) Independent domestication of two old world cotton species. *Genome Biology and Evolution* 8: 1940-1947.