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

Corrinne E Grover[1], Mark A Arick II[1], Justin Conover[1], Adam Thrasher[1], Guanjing Hu[1], William S Sanders[1], Rubab Naqvi[1], Muhammad Farooq[1], Joann Mudge[1], Thiru Ramaraj[1], Joshua A Udall[1], Daniel G Peterson[1], Jodi Scheffler[1], Brian Scheffler[1], Jonathan F Wendel[1]

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

Introduction

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

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

Despite the extensive research on the evolution of Gossypium, these sister genera have been grossly understudied, except in serving as phylogenetic outgroups for cotton phylogenetic and genomic research [10] [11]. Genomic resources in both genera are minimal, access to plant material is limited, and with the recent exception of a study by Sherwood and Morden (2014) on diversity among Kokia species, much of our knowledge regarding these genera is decades old [17] [10] [18].

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

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

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

Materials and Methods

Sequencing and genome assembly

DNA was extracted from mature leaves using the Qiagen Plant DNeasy kit (Qiagen). 350Bp and 550 bp Illumina PCR Free libraries were made and sequenced on 2 Miseq flowcells and 1 Hiseq 2000 lane at the IGBB. The data were trimmed and filtered with Trimmomatic v0.32 [19] with the following options: (1) sequence adapter removal, (2) removal of leading and/or trailing bases when the quality score (Q) ¡28, (3) removal of bases after average Q ¡28 (8 nt window) or single base quality ¡10, and (4) removal of reads ¡85 nt.

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

The trimmed DNA data and RNA assembly were assembled via ABySS v2.0.1 [21], using every 5th kmer value from 65 through 200. The assembly with the highest E-size [22] was used in further analyses. Next the selected assembly was further scaffolded with ABySS using the assembled transcripts. ABySS Sealer v2.0.1 [23] was used to fill gaps in the scaffolded assembly. For each trimmed PCR Free library, Sealer was run with every 10th kmer starting at 100 and decreasing to 30. Pilon v1.22 [24] polished the resulting gap-filled assembly using all the trimmed DNA data.

Genome annotation

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

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

Identification of Orthologs

Amino acid sequences from G. kirkii, G. raimondii and K. drynarioides were clustered using OrthoFinder v1.1.41 [25], which utilizes a Markov clustering algorithm of normalized BLASTp scores to infer homology between proteins sequences of different species. OrthoFinder is similar to OrthoMCL2 [26], but reduces the number of BLAST results by filtering scores based on reciprocal best hits (RBHs) and corrects for gene length biases and floor-limitation of e-values in BLAST scores prior to clustering. These corrections have been shown to increases precision by improved clustering of singletons (i.e., groups in which only one gene from each species is present) instead of entire gene families into a given orthologous group. Default values were used for the inflation parameter (1.5) in the Markov clustering, and the "–og" flag was used to prevent downstream analyses after the groups were generated.

dN/dS Estimation and Timing of Divergence

Singletons inferred from OrthoFinder were separated into all 3 possible pairwise groups (Gr + Gk, Gr + Kd, Kd + Gk). Amino acid sequences from each pairwise group were then aligned using the pairwise2 python package and the BLOSUM62 substitution matrix. The highest scoring alignments were then used as a guide to codon-align the CDS sequences. The CODEML package in PAML [27] was used to calculate the dN, dS, and dN/dS values. Singletons in which any pairwise comparison resulted in a dS value greater than 0.03 was removed from the analysis and inferred to be a cluster of non-orthologous proteins. Distributions of all pairwise dN, dS, and dN/dS values were then plotted, and mean value and standard deviation is reported. Estimates of total divergence time between each pairwise group was calculated using the equation T=dS/(2r) where r is the absolute rate of synonymous substitutions of Adh genes in palms (2.6 X 10-9 substitutions X substitution site-1 X year-1) [11] [28] or members of Brassicaceae (1.5 X 10-8 substitutions X synonymous site-1 X year-1) [29].

Copy Number Variation Estimation

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

Repeat clustering and annotation

All reads from one of the paired-end files (i.e., R1) were filtered for quality and trimmed to a standard 95nt using Trimmomatic version 0.33 [19] as per (https://github.com/Wendellab/KokiaKirkii). Surviving reads were randomly subsampled to represent a 1% genome size equivalent for each genome [30] [31] and combined as input into the RepeatExplorer pipeline [32] [33], 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 [34] using a custom library derived from a combination of Repbase version WHATEVER [35] and previously annotated cotton repeats [36] [37] [38] [39] [40]. A cutoff of 0.01% read representation is common; however, we evaluated the suitability of this cut using a log of diminishing returns (FIGURE WHATEVER; https://github.com/Wendellab/KokiaKirkii).

Within the annotated clusters, the number of megabases (Mb) attributable to that cluster (i.e., element type) for each genome/accession was calculated based on the 1% genome representation of the sample and the standardized read length of 95 nt; total repetitive amounts for each broad repetitive classification were summed from these results. The genome occupation of each cluster (i.e., the calculated number of Mb) was normalized by genome size for each accession, resulting in the percent of each genome occupied by that element type, for use in multivariate visualization (i.e., Principle Coordinate Analysis and Principal Component Analysis). All analyses were conducted in R [41]; 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 [42] and dandelion [43]. Briefly, an all-versus-all BLASTn [44] [**?**] 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 via Bayesian Information Criterion [45]. The read similarity profile for each cluster was automatically evaluated for each histogram to determine if the reads trend toward highly similar "young" or more divergent "older" reads, as per (Julie paper) with an additional category. These categories include (1) positive linear regression; (2) absence of linear regression; (3) negative linear regression; (4) positive quadratic vertical parabola, trend described by right-side of vertex; (4b) positive quadratic vertical parabola, trend described by left-side of vertex; (5) negative quadratic vertical parabola, trend described by right-side of vertex; and (6) negative quadratic vertical parabola, trend described by left-side of vertex and vertex at ¿99% pairwise-identity (Figure WHATEVER). Categories which trend toward highly identical reads (i.e., 1, 4, and 6) were interpreted as having relatively young membership, whereas categories which trend toward lower identity (i.e., 2, 3, 4b, and 5) were interpreted as being composed of older elements. As with Ferreira de Carvalho (2016), this regression simply provides a relative characterization of cluster/element age and is not designed to detect statistically significant differences.

Repetitive profiles between Kokia drynarioides and Gossypioides kirkii

Comparison of abundance for the annotated clusters in Kokia drynarioides and Gossypioides kirkii were visualized via ggplot [46], 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 [47].

Results

Discussion

Supporting Information

S1 Video

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

S1 Text

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

S1 Fig

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

S1 Table

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

Acknowledgments

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

References

1. Dejoode 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 andg. 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.

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. Emms DM, Kelly S (2015) Orthofinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biology 16: 157.

26. Li L, Stoeckert CJ, Roos DS (2003) Orthomcl: Identification of ortholog groups for eukaryotic genomes. Genome Research 13: 2178-2189.

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

28. 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.

29. 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.

30. Hendrix B, Stewart JM (2005) Estimation of the nuclear dna content of gossypium species. Annals of Botany 95: 789-797.

31. Wendel JF, Cronn RC, Spencer Johnston J, James Price H (2002) Feast and famine in plant genomes. Genetica 115: 37-47.

32. 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.

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

34. Smit A, Hubley R, Green P (2013-2015). Repeatmasker open-4.0.

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

36. 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.

37. Grover CE, Hawkins JS, Wendel JF (2008) Phylogenetic insights into the pace and pattern of plant genome size evolution. Genome Dyn 4: 57-68.

38. 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.

39. 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.

40. 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.

41. Team RC (2017). R: A language and environment for statistical computing.

42. 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.

43. 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.

44. 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.

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

46. Wickham H (2016) ggplot2: elegant graphics for data analysis. Springer.

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