Re-evaluating the phylogeny of allopolyploid *Gossypium* L

Corrinne E Grover¹, Joseph P Gallagher¹, Josef J Jareczek¹, Justin T Page², Joshua A Udall², Michael A. Gore³, and Jonathan F Wendel¹

Affiliations

¹Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA 50011

²Plant and Wildlife Science Department, Brigham Young University, Provo, UT 84602

³ Plant Breeding and Genetics Section, School of Integrative Plant Science, Cornell University, Ithaca, NY 14853

Corresponding author info

Corrinne Grover; 251 Bessey Hall, Department of Ecology, Evolution and Organismal Biology, Iowa State University, Ames IA 50011; phone: 515-294-7499; Fax: 515-294-1337; corrinne@iastate.edu

Abstract

The formation of allopolyploid cotton precipitated a rapid diversification and colonization of dry coastal American tropical and subtropical regions. Previous phylogenetic analyses, combined with molecular divergence analyses, have offered a temporal framework for this radiation, but provide only weak support for some of the resolved branches. Moreover, these earlier analyses did not include the recently recognized sixth polyploid species, *G. ekmanianum* Wittmack. Here we use targeted sequence capture of multiple loci in conjunction with both concatenated and Bayesian concordance analyses to reevaluate the phylogeny of allopolyploid cotton species. Although phylogenetic resolution afforded by individual genes is often low, sufficient signal was attained both through the concatenated and concordance analyses to provide robust support for the *Gossypium* polyploid clade, which is reported here.

Keywords: cotton, *Gossypium*, allopolyploid, phylogeny, sequence capture

1. Introduction

Gossypium is a remarkably diverse genus, with over 50 recognized species divided into 8 diploid genome groups (Endrizzi et al., 1985; Wendel and Grover, 2015) and a single, monophyletic (Grover et al., 2012) allopolyploid lineage. Also known as the cotton genus, *Gossypium* contains the world's largest source of renewable plant-based textile fiber. Although the allopolyploid species *G. hirsutum* L., or upland cotton, presently dominates global cotton commerce, historically there have been independent domestications of four different cotton species, two allopolyploid and two diploid. The other domesticated allopolyploid, *G. barbadense* L. (or Pima cotton), is prized for its long, strong fibers (<http://www.cotton.org>). Most commercially grown cotton (~95%) derives from these two allopolyploid species, drawing considerable attention to the relationships among these two domesticated and the other four recognized polyploid species, *G. tomentosum* Nutt ex Seem, *G. mustelinum* Miers ex Watt, *G. darwinii* Watt, and *G. ekmanianum* (Grover et al., 2014).

While the genus itself has a relatively broad aggregate distribution in the arid to sub-arid tropics and subtropics (Wendel and Cronn, 2003; Wendel and Grover, 2015), the native distribution of allopolyploid cotton is much narrower. A remarkable transoceanic dispersal of an African, A-genome diploid to the New World led to the founding of the nascent allopolyploid species when the newly arrived species merged with a native D-genome diploid and subsequently underwent genome doubling, approximately 1-2 million years ago. This was followed by rapid radiation of the new AD-genome clade into the coastal tropical and subtropical regions of the Caribbean, Central America, northern South America, and even dispersal into more distant island chains (e.g., The Galapagos, Hawaii and the south Pacific).

The geographic spread of the allopolyploid lineage was accompanied by rapid diversification and speciation, which has hindered confident reconstruction of their phylogeny. Many early efforts involving multiple molecular markers (e.g., cpDNA restriction site analysis, rDNA ITS sequencing, nuclear and chloroplast genes) obtained weak or limited resolution (Cronn et al., 1996; DeJode and Wendel, 1992; Endrizzi et al., 1985; Reinisch et al., 1994; Seelanan et al., 1997; Small et al., 1998; Wendel, 1989; Wendel and Albert, 1992; Wendel et al., 1995a; Wendel et al., 1995b). Two phylogenetic inferences appear to be robust, including the basal split between a clade now represented by the Brazilian endemic *G. mustelinum* Miers ex Watt and the remaining species and the sister species status of *G. barbadense* and *G. darwinii* Watt (these once considered conspecific; (Wendel and Percy, 1990)). Nonetheless, and in spite of these extensive efforts, the branching order of several nodes has remained unclear.

Rapid radiations present a particular challenge for phylogenetic reconstructions, due to several phenomena, including incomplete lineage sorting and low signal to noise ratio, the latter leading to spurious homoplasies that can confound inference of historical branching pattern (Avice and Wollenberg, 1997; Boussau et al., 2013; Degnan and Rosenberg, 2009; Eaton and Ree, 2013; Mirarab et al., 2014a; Pamilo and Nei, 1988; Rokas et al., 2003; Szöllősi and Daubin, 2012; Szöllősi et al., 2015). A common approach to overcome this challenge is to use multiple individual loci (Barrett et al., 1991; de Queiroz and Gatesy, 2007; Degnan and Rosenberg,

2006; Degnan and Rosenberg, 2009; Huelsenbeck et al., 1996; Kubatko and Degnan, 2007; Mirarab et al., 2014a; Song et al., 2012; Wielstra et al., 2014), and, with the advent of high-throughput genomic technologies, even full phylogenomic data sets consisting of hundreds to thousands of markers distributed throughout the genome (Arbizu et al., 2014; Bond et al., 2014; Boussau et al., 2013; Brewer and Bond, 2013; Dunn et al., 2013; Eaton and Ree, 2013; Johnson et al., 2013; Jones, 2015; Leigh et al., 2011; Pecon-Slattery, 2014; Song et al., 2012; Stephens et al., 2015; Wen et al., 2013; Zimmer and Wen, 2012). Here we combine targeted sequence capture with next-generation sequencing in an effort to more confidently resolve the phylogeny of allopolyploid cotton, and to phylogenetically place the recently resurrected species, *G. ekmanianum* (Grover et al., 2014; Krapovickas and Seijo, 2008).

2. Materials and Methods

2.1 Plant materials, sequence capture, and sequencing. DNA was extracted from fresh leaves of: (1) wild and domesticated *G. hirsutum* and *G. barbadense*; (2) wild accessions of the remaining four allopolyploid species, *G. tomentosum*, *G. mustelinum*, *G. darwinii*, and *G. ekmanianum*; and (3) one representative accession each for the model A-genome and D-genome diploid parents, *G. arboreum* and *G. raimondii*, respectively. In total, 24 accessions of allopolyploids were included, representing eight each of *G. hirsutum* and *G. barbadense*, three *G. tomentosum* Nutt ex Seem, two each of *G. mustelinum* and *G. darwinii*, and one *G. ekmanianum* (Table 1). Targeted sequence capture was performed on the selected *G. hirsutum* accessions, *G. arboreum* L. (A2), and *G. raimondii* Ulbrich (D5), using NimbleGen slide captures according to the protocol described by Salmon et al (2012), using a probe set targeting approximately 500 genes. These genes were originally selected from the literature and from expression experiments based on their potential role in cotton fiber development; a full list is at <http://cottonrevolution.info>, and in Salmon et al (2012). Targeted sequence capture of the *G. barbadense* accessions and the remaining polyploid accessions was performed using the Mycroarray MYbaits custom bait libraries according to the manufacturer's protocol (<http://www.mycroarray.com/pdf/MYbaits-manual.pdf>), with an expanded target set that includes the original 500 genes targeted above, as well as additional genes representing members of the anthocyanin biosynthetic pathway and flowering time network. Custom baits targeting these genes were designed and synthesized by Mycroarray, and library construction was completed using the Roche Rapid Library preparation protocol at the Georgia Genomics Facility (<http://dna.uga.edu/>) using rapid library multiplex identifiers 1-12 (RL 1-12). Sequence capture and post-capture amplification were performed according to the manufacturer's protocol.

For both sequence capture data sets, quantitative PCR (qPCR) enrichment verification was used, as previously reported (Salmon et al., 2012), and both the quantity and quality of the captured libraries was assessed using the Bioanalyzer DNA7500 chip (Agilent). Libraries were pooled in sets of 12, and each pool of 12 was sequenced on a GS-FLX Titanium 454 Sequencing System (DNA Sequence Center, Brigham Young University, Provo, UT).

2.2 Sequence assembly and alignment. All reads were trimmed using sickle (Joshi and Fass, 2011) with default settings and subsequently mapped onto the recently published *G. raimondii* genome (Paterson et al., 2012) using GSNAP (Wu and Nacu, 2010) in combination with species-specific SNP indices (Page et al., 2013a). For the allopolyploids, reads arising from each parental genome were partitioned into those arising from each homoeolog using PolyCat (Page et al., 2013a) in conjunction with the specific SNP index used during mapping. Gene consensus sequences and alignments were generated from the mapped and segregated bam files using bam2consensus from the BamBam software package (Page et al., 2014), which returned consensus sequences for >37,000 genes predicted in the *G. raimondii* genome (Paterson et al., 2012).

2.3 Data filtering. Because allopolyploid genomes often are subject to gene conversion (Flagel et al., 2012; Guo et al., 2014; Salmon et al., 2010; Wang and Paterson, 2011), we first filtered the data using a strict approach implemented using the ape package (Paradis et al., 2004) of R (R Core Team, 2013). Briefly, this script excluded any gene where diagnostic SNPs (based on the diploid orthologs) were not faithfully reproduced in the two homoeologous allopolyploid reads. Subsequently, the remaining genes were ranked by read coverage, as output by counter (Page et al., 2013b). Fifty-two genes with the highest read coverage (i.e., more than 80% of the accessions had a read count >25 and more than half had a read count >50) were selected for subsequent analyses (Table 2). For these alignments, positions that contained more than 25% Ns were removed using a custom python script based on PyCogent (Knight et al., 2007). All scripts pertaining to this project are at <https://github.com/Wendellab/phylogenetics>.

2.4 Phylogenetic and concordance analysis. Phylogenetic relationships were inferred using Neighbor-Joining, Maximum Likelihood, and Bayesian analyses of individual genes and also using a concatenated alignment. Neighbor-Joining trees were generated using ClustalW (Larkin et al., 2007) and excluded gapped regions. Maximum likelihood analyses were performed using RaxML (Stamatakis, 2014) using the basic general time reversible model with gamma distribution (GTRGAMMA), 1000 alternative runs on distinct starting trees, and rapid bootstrapping with consensus generation. Bayesian analyses were generated using MrBayes (Ronquist and Huelsenbeck, 2003) under GTR gamma with the following parameters: 3 runs with four chains for 10 million generations and using a burn-in fraction of 25%. Phylogenetic analyses were conducted for the A- and D-homoeologs independently, using members of each respective diploid genome group to root the resulting trees.

Congruence among individual gene trees was evaluated by three methods. In the first method, topologies were assessed to fit the generalized polyploid gene trees (Supplementary Figures). These were tabulated for the individual Neighbor-Joining and Maximum Likelihood. A formal estimate of concordance was made for the individual Bayesian trees via BUCKy (Ané et al., 2007; Larget et al., 2010) using default parameters, 3 runs each with 4 chains, and 1 million iterations. Since BUCKy requires common taxa for all gene trees, a basic set of taxa was independently selected for both the A and D gene sets using a single representative accession for each species that included sequence for most genes in the common set of 52 genes evaluated. Despite restricting the accessions to those with the highest gene coverage, 14-18

genes (A and D, respectively) were excluded by the program due to missing data, 11 of which were common to both A and D. A list of genes and taxa used for concordance analyses is presented in Tables 1 and 2. Finally, a coalescence-based estimation of the species tree was derived for each homoeologous gene set using ASTRAL (Mirarab et al., 2014b).

The biogeographical radiation of the polyploid clade was formally addressed using RASP (Yu et al., 2015). Output is in supplemental figures.

3. Results and Discussion

3.1 Rapid divergence and low per gene resolution in individual gene phylogenies

As previously observed, the rapid and recent radiation of polyploid cotton hinders a full resolution of species relationships on a per gene basis. For the 38 genes analyzed from the A genome of the allopolyploids (the A_T homoeologs), 22 have 100% sequence identity between two or more species or accessions, contributing to the relatively high average maximum sequence identity within each homoeologous alignment of 99.95%. This is perhaps unsurprising, as multiple accessions were included for each species; however, the observed distance between the most divergent sequences was also low, with the most divergent sequences within the A_T homoeolog gene sets still retaining 98.29% similarity (average minimum pairwise divergence across all genes = 99.35%). The D_T homoeolog gene sets were similarly identical. This low level of sequence divergence observed for each gene is also reflected in the lack of resolution among species for each gene tree (Supplementary Figures).

The geographic radiation of the allopolyploid lineage was also tested here using both Bayesian Binary MCMC and Bayesian inference for discrete areas (BayArea), as implemented in RASP (Yu et al., 2015). Both methods produce similar results, in that each species has a distinct, and often isolated, geographical area with the ancestral nodes representing most of the areas (Supplementary Figure). This is unsurprising, given the propensity for cotton to exist in small, isolated populations and the observation that many of these species exist as island endemics (e.g. *G. tomentosum* in Hawaii and *G. darwinii* in the Galapagos) (Wendel and Grover, 2015).

3.2 Resolving the phylogeny of polyploid *Gossypium*

Because the high similarity within individual genes provided little phylogenetic resolution, we concatenated all genes for a given accession, in an effort to boost phylogenetic signal. For each combined dataset, one for each set of homoeologs (A_T and D_T), we generated trees via NJ, ML, and Bayesian analyses (Figure 1). For all analyses, two sister taxa relationships were recovered, i.e., between *G. barbadense* and *G. darwinii* (AD2, AD5) and between *G. hirsutum* and *G. ekmanianum* (AD1, AD6). The first of these mirrors all earlier results, whereas the latter is a novel result of the present study. NJ and ML analyses for the same homoeologous data set yielded either identical (Fig. 1, panels A, B, for the A_T dataset) or nearly identical (Fig. 1, panels D, E, for the D_T dataset) topologies, but these two sets of topologies resolved different trees (Figure 1). The A-homoeologs resolved the historically accepted phylogeny, where *G. mustelinum* (AD4) represents the basal-most lineage, followed by radiations of the *G. tomentosum* – *G. hirsutum* (AD3, AD1) and *G. barbadense* – *G. darwinii* (AD2, AD5) lineages.

The D-homoeologs, however, differed in placing the sister clade of *G. barbadense* – *G. darwinii* (AD2, AD5) as the basal-most lineage, followed by branching of each of the remaining polyploids (in differing orders; Figure 1). In addition, the Bayesian analyses provide weak but different resolutions for both concatenated datasets (Figure 1, panels C, F). When we concatenated all homoeologs into a single dataset, however, the topology resolved in each case (NJ, ML, and Bayesian) perfectly mirrors the topology resolved for the A-homoeolog individual dataset, despite equivalent representation of A- and D-homoeologs in the data (86 kb and 85 kb of A_T and D_T, respectively). The discrepancies between the resolutions provided by the two homoeologous concatenated data sets, which must have the same species history, can be explained by undetected gene conversion or paralogy, or perhaps other phenomena that cause gene trees to deviate from species history, particularly, we suspect, the occurrence of spurious homoplasy that overwhelms the limited phylogenetic signal that exists because of rapid cladogenesis.

One criticism of using concatenated trees is that by merging datasets, the ability to detect conflict among gene trees is removed. Evaluating conflict among gene trees can be informative in determining sources of discordance, and focus attention on the set of genes that is responsible for the differing resolutions. This conflict among branching patterns inferred from individual genes can be observed and quantified. In our case, all data support the two sister-species pairs AD1/AD6 and AD2/AD5. Thus, we have four representative terminals (AD1/AD6, AD2/AD5, AD3, and AD4) that may resolve into only a limited number of topologies (Figure S1). Considering this, we assigned individual gene trees from the NJ and ML analyses to one of the topologies outlined and quantified the most frequent topologies observed (Figure 2). For both NJ and ML analyses, the most frequently observed topology was a polytomy among the four representative terminals (Figure 2, left panel). This is unsurprising, given the high sequence identity and lack of resolution provided by most genes, as noted above. The next most frequently observed tree differed for the NJ and ML analyses, with resolution of the historically accepted phylogeny being most frequent for NJ (Figure 2, center panel) and a tie between two less resolved phylogenies being most frequent for ML (Figure 2, left and center). Overall, most genes supported or did not positively conflict with the historically hypothesized phylogeny (Figure 2, top center), which places the Hawaiian Islands endemic *G. tomentosum* (AD3) as sister to *G. hirsutum* (AD1). Support for this earlier hypothesis stems largely from a single nuclear gene, with little to no support from chloroplast restriction site analyses. Here, we note that the vast majority of gene trees shown in Figure 2 either positively support this sister-species relationship (Figure 2 center panel, representing 22 NJ and 28 ML trees, respectively) or remain agnostic in this respect (Figure 2 left panel, representing 44 and 53 NJ and ML trees, respectively), with only a minority fraction of trees supporting conflicting resolutions (Figure 2 right panel, representing 10 and 8 NJ and ML trees, respectively).

To more formally evaluate the consensus among trees, we employed Bayesian concordance analysis via BUCKy (Ané et al., 2007; Larget et al., 2010). Bayesian concordance seeks to estimate the proportion of genes that reflect the true species tree (represented by the most frequently resolved clades) versus those in conflict (Baum, 2007), providing concordance factors for individual clades. For this analysis, we evaluated concordance among the A- and D-homoeologs separately. The A-homoeologs generally had high concordance in resolving the

phylogeny of the cotton allopolyploids (Figure 3). The concordance analysis identified five splits (Figure 3), ranging in CF from a low of 0.106 for the *G. barbadense* - *G. darwinii* - *G. tomentosum* clade (AD2, AD5, AD3) to a high of 0.95 for the *G. hirsutum* and *G. ekmanianum* (AD1, AD6), providing robust support for this sister-species relationship. High CF values were also obtained for the *G. barbadense* - *G. darwinii* relationship (0.93) and the basal status for *G. mustelinum* (0.91), as well as for the union of AD1/AD6 with AD3 (CF = 0.84). The alternative relationship for AD3, i.e., with AD2/AD5, had a far lower CF (0.11) than all other splits observed and represented the only conflicting split in the A-homoeolog data set (Figure 3). Overall, the A-homoeologs overwhelmingly support the placement of *G. tomentosum* as sister to the *G. hirsutum* - *G. ekmanianum* (AD1, AD6) clade, in keeping with both the bulk of the individual gene tree analysis discussed above and the historically hypothesized phylogeny (e.g., Wendel and Cronn, 2003).

Concordance analysis of the D-homoeologs, however, resolved more splits with lower average CFs, indicating more discordance in these genes. As with the A-homoeologs, the top four splits for the D-homoeologs resolved *G. tomentosum* as sister to the *G. hirsutum* - *G. ekmanianum* (AD1, AD6) clade, which was also resolved with high CF (0.70 and 0.78, respectively; Figure 3). The confidence intervals for these four splits did not overlap with the remaining conflicting splits, the topmost conflicting split again placing *G. tomentosum* sister to *G. barbadense* and *G. darwinii*. Despite the discordance in the trees, both homoeologs produced the primary concordance tree depicted in Figure 3A, which, given that none of the conflicting splits from either homoeologous set overlapped in CI with the splits, largely supports that phylogeny as the best-supported species tree.

The phylogenetic tree with the highest concordance (Figure 3, Tree A) was also recovered in a separate coalescent analysis for both homoeologous gene sets (trees not shown), providing further support for the validity of these relationships. At present, we have no clear explanation for the conflict evident in one homoeologous set (D_T) and not present in the other (A_T). Gene conversion, which may experience directional biases (Flagel et al., 2012; Guo et al., 2014; Salmon et al., 2010), is unlikely to have a large effect given our strict filtering parameters. While minimized in the present dataset, the amount of missing data or sequencing error could contribute noise to the tree; however, this is unlikely to be biased toward one homoeologous genome. Likewise, regions of possible alignment errors were excluded, and are unlikely to produce the pattern seen here. Given the rapid diversification of the polyploid species, lack of coalescence and incomplete lineage sorting are reasonable potential sources of conflict. Cryptic paralogy (copy number variation) may also create conflict among trees, which becomes reasonable in light of recent data (Page et al, in prep) which suggests that the rate and fixation of large indels (including genes) is different between the A_T and D_T homoeologous genomes and among species.

4. Conclusions

Accurate phylogenies provide the necessary framework for understanding historical biogeography and evolutionary change among species. Here, using a targeted sequence

capture approach and high throughput sequencing, we reevaluated previously hypothesized relationships among *Gossypium* allopolyploids, some of which were only weakly supported, and in addition including a newly named sixth allopolyploid species, *G. ekmanianum*. The use of multiple loci analyzed from multiple perspectives, including differing phylogenetic approaches and single genes vs. concatenated data sets, combined with statistical concordance analyses, allowed us to provide a more robust phylogeny for allopolyploid *Gossypium* (summarized in Figure 4). Key conclusions are that the most recently named allopolyploid species, *G. ekmanianum*, is sister to *G. hirsutum*; that these two species form a clade that is sister to the Hawaiian Island endemic *G. tomentosum*; that the Galapagos Islands endemic *G. darwinii* is sister to *G. barbadense*; that this latter pair of species is sister to the *G. ekmanianum*-*G. hirsutum*-*G. tomentosum* clade; and that *G. mustelinum* is isolated as one branch of the earliest split following allopolyploid formation (Figure 3). These relationships appear robust, and moreover, they implicate a biogeographic scenario of rapid radiation and salt water dispersal throughout the American tropics and subtropics. In this sense *Gossypium* may exemplify novel ecological opportunity afforded by allopolyploidy, in that in contrast to most diploids in the genus, allopolyploid *Gossypium* typically are found in coastal beach strand vegetation, even occurring in active salt-spray zones or other coastal areas subject to flooding. As noted earlier (Fryxell, 1978; Stephens, 1958; Wendel, 1989), this ecological preference may be causally connected to their rapid colonization and spread throughout the New World tropics.

Acknowledgments

We thank Richard Percy for assistance in selecting the *G. barbadense* accessions, Emma Dostal for help with evaluating the trees, and David Baum for helpful phylogenetic discussions. We acknowledge support for this project from the Plant Genome Program of the National Science Foundation, USDA-ARS, and Cotton Inc. Joseph P. Gallagher is funded by a Graduate Research Fellowship from the National Science Foundation.

References

- Ané, C., Larget, B., Baum, D.A., Smith, S.D., Rokas, A., 2007. Bayesian Estimation of Concordance among Gene Trees. *Molecular Biology and Evolution* 24, 412-426.
- Arbizu, C., Ruess, H., Senalik, D., Simon, P.W., Spooner, D.M., 2014. Phylogenomics of the carrot genus (*Daucus*, Apiaceae). *American Journal of Botany*.
- Avice, J.C., Wollenberg, K., 1997. Phylogenetics and the origin of species. *Proceedings of the National Academy of Sciences of the United States of America* 94, 7748-7755.
- Barrett, M., Donoghue, M.J., Sober, E., 1991. Against consensus. *Systematic Zoology*, 486-493.
- Baum, D.A., 2007. Concordance trees, concordance factors, and the exploration of reticulate genealogy. *Taxon* 56, 417-426.
- Bond, Jason E., Garrison, Nicole L., Hamilton, Chris A., Godwin, Rebecca L., Hedin, M., Agnarsson, I., 2014. Phylogenomics Resolves a Spider Backbone Phylogeny and Rejects a Prevailing Paradigm for Orb Web Evolution. *Current Biology* 24, 1765-1771.
- Boussau, B., Szöllösi, G.J., Duret, L., Gouy, M., Tannier, E., Daubin, V., 2013. Genome-scale coestimation of species and gene trees. *Genome Res* 23, 323-330.
- Brewer, M.S., Bond, J.E., 2013. Ordinal-Level Phylogenomics of the Arthropod Class Diplopoda (Millipedes) Based on an Analysis of 221 Nuclear Protein-Coding Loci Generated Using Next-Generation Sequence Analyses. *PLoS ONE* 8, e79935.
- Cronn, R., Zhao, X., Paterson, A., Wendell, J., 1996. Polymorphism and concerted evolution in a tandemly repeated gene family: 5S ribosomal DNA in diploid and allopolyploid cottons. *J Mol Evol* 42, 685-705.
- de Queiroz, A., Gatesy, J., 2007. The supermatrix approach to systematics. *Trends in Ecology & Evolution* 22, 34-41.
- Degnan, J.H., Rosenberg, N.A., 2006. Discordance of species trees with their most likely gene trees. *PLoS genetics* 2, e68.
- Degnan, J.H., Rosenberg, N.A., 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in ecology & evolution* 24, 332-340.
- DeJode, D.R., Wendel, J.F., 1992. Genetic Diversity and Origin of the Hawaiian Islands Cotton, *Gossypium tomentosum*. *American Journal of Botany* 79, 1311-1319.
- Dunn, C., Howison, M., Zapata, F., 2013. Agalma: an automated phylogenomics workflow. *BMC Bioinformatics* 14, 330.
- Eaton, D.A.R., Ree, R.H., 2013. Inferring Phylogeny and Introgression using RADseq Data: An Example from Flowering Plants (*Pedicularis*: Orobanchaceae). *Systematic Biology* 62, 689-706.
- Endrizzi, J.D., Turcotte, E.L., Kohel, R.J., 1985. Genetics, cytology, and evolution of *Gossypium*. *Advances in Genetics* 23, 271-375.
- Flagel, L., Wendel, J., Udall, J., 2012. Duplicate gene evolution, homoeologous recombination, and transcriptome characterization in allopolyploid cotton. *BMC Genomics* 13, 302.
- Fryxell, P.A., 1978. The natural history of the cotton tribe (Malvaceae, tribe Gossypieae). Texas A & M University Press.
- Grover, C.E., Grupp, K.K., Wanzek, R.J., Wendel, J.F., 2012. Assessing the monophyly of polyploid *Gossypium* species. *Plant Syst Evol* 298, 1177-1183.
- Grover, C.E., Zhu, X., Grupp, K.K., Jareczek, J.J., Gallagher, J.P., Szadkowski, E., Seijo, J.G., Wendel, J.F., 2014. Molecular confirmation of species status for the allopolyploid cotton species, *Gossypium ekmanianum* Wittmack. *Genet Resour Crop Evol*, 1-12.

Guo, H., Wang, X., Gundlach, H., Mayer, K.F.X., Peterson, D.G., Scheffler, B.E., Chee, P.W., Paterson, A.H., 2014. Extensive and Biased Intergenomic Nonreciprocal DNA Exchanges Shaped a Nascent Polyploid Genome, *Gossypium* (Cotton). *Genetics* 197, 1153-1163.

Huelsenbeck, J.P., Bull, J.J., Cunningham, C.W., 1996. Combining data in phylogenetic analysis. *Trends in Ecology & Evolution* 11, 152-158.

Johnson, Brian R., Borowiec, Marek L., Chiu, Joanna C., Lee, Ernest K., Atallah, J., Ward, Philip S., 2013. Phylogenomics Resolves Evolutionary Relationships among Ants, Bees, and Wasps. *Current Biology* 23, 2058-2062.

Jones, B., 2015. Phylogenomics: Building the insect tree-of-life. *Nature Reviews Genetics* 16, 2-3.

Joshi, N.A., Fass, J.N., 2011. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software].

Knight, R., Maxwell, P., Birmingham, A., Carnes, J., Caporaso, J.G., Easton, B., Eaton, M., Hamady, M., Lindsay, H., Liu, Z., Lozupone, C., McDonald, D., Robeson, M., Sammut, R., Smit, S., Wakefield, M., Widmann, J., Wikman, S., Wilson, S., Ying, H., Huttley, G., 2007. PyCogent: a toolkit for making sense from sequence. *Genome Biology* 8, R171.

Krapovickas, A., Seijo, G., 2008. *Gossypium ekmanianum* (malvaceae), algodón silvestre de la República Dominicana. *Bonplandia* 17, 55-63.

Kubatko, L.S., Degnan, J.H., 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst Biol* 56, 17-24.

Larget, B.R., Kotha, S.K., Dewey, C.N., Ané, C., 2010. BUCKy: Gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics* 26, 2910-2911.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G., 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 2947-2948.

Leigh, J.W., Lapointe, F.-J., Lopez, P., Baptiste, E., 2011. Evaluating Phylogenetic Congruence in the Post-Genomic Era. *Genome Biology and Evolution* 3, 571-587.

Mirarab, S., Bayzid, M.S., Warnow, T., 2014a. Evaluating Summary Methods for Multilocus Species Tree Estimation in the Presence of Incomplete Lineage Sorting. *Systematic Biology*.

Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S., Warnow, T., 2014b. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* 30, i541-i548.

Page, J.T., Gingle, A.R., Udall, J.A., 2013a. PolyCat: A Resource for Genome Categorization of Sequencing Reads From Allopolyploid Organisms. *G3-Genes Genom Genet* 3, 517-525.

Page, J.T., Huynh, M.D., Liechty, Z.S., Grupp, K., Stelly, D., Hulse, A.M., Ashrafi, H., Van Deynze, A., Wendel, J.F., Udall, J.A., 2013b. Insights into the evolution of cotton diploids and polyploids from whole-genome re-sequencing. *G3* 3, 1809-1818.

Page, J.T., Liechty, Z.S., Huynh, M.D., Udall, J.A., 2014. BamBam: genome sequence analysis tools for biologists. *BMC Research Notes* 7, 829.

Pamilo, P., Nei, M., 1988. Relationships between gene trees and species trees. *Molecular biology and evolution* 5, 568-583.

Paradis, E., Claude, J., Strimmer, K., 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20, 289-290.

Paterson, A.H., Wendel, J.F., Gundlach, H., Guo, H., Jenkins, J., Jin, D., Llewellyn, D., Showmaker, K.C., Shu, S., Udall, J., Yoo, M.-j., Byers, R., Chen, W., Doron-Faigenboim, A., Duke, M.V., Gong, L., Grimwood, J., Grover, C., Grupp, K., Hu, G., Lee, T.-h., Li, J., Lin, L., Liu, T., Marler, B.S., Page, J.T., Roberts, A.W., Romanel, E., Sanders, W.S., Szadkowski, E., Tan, X., Tang, H., Xu, C., Wang, J., Wang, Z., Zhang, D., Zhang, L., Ashrafi, H., Bedon, F., Bowers, J.E., Brubaker, C.L., Chee, P.W., Das, S., Gingle, A.R., Haigler, C.H., Harker, D., Hoffmann, L.V., Hovav, R., Jones, D.C., Lemke, C., Mansoor, S., Rahman, M.u., Rainville, L.N., Rambani, A., Reddy, U.K., Rong, J.-k., Saranga, Y., Scheffler, B.E., Scheffler, J.A., Stelly, D.M., Triplett, B.A., Van Deynze, A., Vaslin, M.F.S., Waghmare, V.N., Walford, S.A., Wright, R.J., Zaki, E.A.,

Zhang, T., Dennis, E.S., Mayer, K.F.X., Peterson, D.G., Rokhsar, D.S., Wang, X., Schmutz, J., 2012. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature* 492, 423-427.

Pecon-Slaterry, J., 2014. Recent Advances in Primate Phylogenomics. *Annual Review of Animal Biosciences* 2, 41-63.

R Core Team, 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Reinisch, A.J., Dong, J.M., Brubaker, C.L., Stelly, D.M., Wendel, J.F., Paterson, A.H., 1994. A detailed RFLP map of cotton, *Gossypium hirsutum* x *Gossypium barbadense*: chromosome organization and evolution in a disomic polyploid genome. *Genetics* 138, 829-847.

Rokas, A., Williams, B.L., King, N., Carroll, S.B., 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425, 798-804.

Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572-1574.

Salmon, A., Flagel, L., Ying, B., Udall, J.A., Wendel, J.F., 2010. Homoeologous nonreciprocal recombination in polyploid cotton. *New Phytologist* 186, 123-134.

Salmon, A., Udall, J.A., Jeddelloh, J.A., Wendel, J., 2012. Targeted Capture of Homoeologous Coding and Noncoding Sequence in Polyploid Cotton. *G3-Genes Genom Genet* 2, 921-930.

Seelanan, T., Schnabel, A., Wendel, J.F., 1997. Congruence and consensus in the cotton tribe. *Systematic Botany* 22, 259-290.

Small, R.L., Ryburn, J.A., Cronn, R.C., Seelanan, T., Wendel, J.F., 1998. The tortoise and the hare: choosing between noncoding plastome and nuclear *Adh* sequences for phylogeny reconstruction in a recently diverged plant group. *American Journal of Botany* 85, 1301-1315.

Song, S., Liu, L., Edwards, S.V., Wu, S., 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proceedings of the National Academy of Sciences*, 201211733.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312-1313.

Stephens, J.D., Rogers, W.L., Heyduk, K., Cruse-Sanders, J.M., Determann, R.O., Glenn, T.C., Malmberg, R.L., 2015. Resolving phylogenetic relationships of the recently radiated carnivorous plant genus *Sarracenia* using target enrichment. *Molecular Phylogenetics and Evolution* 85, 76-87.

Stephens, S.G., 1958. Salt Water Tolerance of Seeds of *Gossypium* Species as a Possible Factor in Seed Dispersal. *The American Naturalist* 92, 83-92.

Szöllősi, G.J., Daubin, V., 2012. Modeling gene family evolution and reconciling phylogenetic discord. *Evolutionary Genomics*. Springer, pp. 29-51.

Szöllősi, G.J., Tannier, E., Daubin, V., Boussau, B., 2015. The Inference of Gene Trees with Species Trees. *Systematic Biology* 64, e42-e62.

Wang, X.-Y., Paterson, A.H., 2011. Gene Conversion in Angiosperm Genomes with an Emphasis on Genes Duplicated by Polyploidization. *Genes* 2, 1-20.

Wen, J., Xiong, Z., Nie, Z.-L., Mao, L., Zhu, Y., Kan, X.-Z., Ickert-Bond, S.M., Gerrath, J., Zimmer, E.A., Fang, X.-D., 2013. Transcriptome Sequences Resolve Deep Relationships of the Grape Family. *PLoS ONE* 8, e74394.

Wendel, J.F., 1989. New World tetraploid cottons contain Old World cytoplasm. *Proceedings of the National Academy of Sciences* 86, 4132-4136.

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

Wendel, J.F., Cronn, R.C., 2003. Polyploidy and the evolutionary history of cotton. *Advances in Agronomy* 78, 139-186.

- Wendel, J.F., Grover, C.E., 2015. Taxonomy and evolution of the cotton genus, *Gossypium*. In: Fang, D.D., Percy, R.G. (Eds.), Cotton, 2nd edition.
- Wendel, J.F., Percy, R.G., 1990. Allozyme diversity and introgression in the Galapagos Islands endemic *Gossypium darwinii* and its relationship to continental *G. barbadense*. *Biochemical Systematics and Ecology* 18, 517-528.
- Wendel, J.F., Schnabel, A., Seelanan, T., 1995a. Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*). *Proceedings of the National Academy of Sciences* 92, 280-284.
- Wendel, J.F., Schnabel, A., Seelanan, T., 1995b. An Unusual Ribosomal DNA Sequence from *Gossypium gossypoides* Reveals Ancient, Cryptic, Intergenomic Introgression. *Molecular Phylogenetics and Evolution* 4, 298-313.
- Wielstra, B., Arntzen, J.W., van der Gaag, K.J., Pabijan, M., Babik, W., 2014. Data Concatenation, Bayesian Concordance and Coalescent-Based Analyses of the Species Tree for the Rapid Radiation of *Triturus* Newts. *PLoS ONE* 9, e111011.
- Wu, T.D., Nacu, S., 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 26, 873-881.
- Yu, Y., Harris, A.J., Blair, C., He, X., 2015. RASP (Reconstruct Ancestral State in Phylogenies): A tool for historical biogeography. *Molecular Phylogenetics and Evolution* 87, 46-49.
- Zimmer, E.A., Wen, J., 2012. Using nuclear gene data for plant phylogenetics: Progress and prospects. *Molecular Phylogenetics and Evolution* 65, 774-785.

Figure legends:

Figure 1. Concatenated trees for the A- and D- homoeologs using Neighbor-Joining (NJ), Maximum Likelihood (ML), and Bayesian (MB) analyses. Panels A – C represent the concatenated trees resolved by the concatenated A-homoeologs using NJ, ML, and MB reconstructions, respectively (see methods), while Panels D – E represent the concatenated D-homoeologs in the same order. The A- and D- homoeolog-derived trees were consistent within homoeologs, irrespective of the method of reconstruction; however, the MB analyses typically provided lower overall resolution. Concatenation of all homoeologs together resolved trees congruent with panels A – C.

Figure 2. Most frequently observed topologies in the individual NJ and ML trees. The most frequently observed topologies are depicted, with the observed number of occurrences listed below each topology. Blue corresponds to NJ, while green corresponds to ML. The left most panel represents the unresolved relationships among the polyploid clades, the center panel represents those topologies that support *G. tomentosum* as sister to *G. ekmanianum*-*G. hirsutum* whereas the right panel depicts the topologies that support the alternate, *i.e.*, *G. tomentosum* sister to *G. barbadense*-*G. darwinii*.

Figure 3. Alternate trees suggested by Bayesian Concordance Analysis. Two trees are depicted, representing (A) the primary concordance tree for both A- and D- homoeologous gene sets (green), and (B) the tree representing the second most supported splits (blue). For each homoeologous set, the numbered splits are listed, including their concordance factors (CF), confidence intervals (CI), and the taxa that the split unites. For the D-homoeologous set, multiple splits with lower concordance factors were detected (shown in grey), and split [6] is highlighted in both blue and grey to indicate that, while it has a lower CF than split [5], their confidence intervals overlap. The taxa included are *G. hirsutum* (AD1), *G. barbadense* (AD2), *G. tomentosum* (AD3), *G. mustelinum* (AD4), *G. darwinii* (AD5), and *G. ekmanianum* (AD6). Each primary concordance tree was rooted with the appropriate diploid, *G. arboreum* (A2) for the A-homoeologous set (pictured), or *G. raimondii* (D5) for the D-homoeologous set (not pictured).

Figure 4. Summary phylogeny for polyploid *Gossypium*. The best-supported polyploid phylogeny is depicted here, which is the result of multiple analyses that gave largely congruent topologies. ML-derived branch lengths are shown.

Figure 1

Figure 2

Figure 3

A homoeologs

	CF	CI	
[1]	0.950	(0.868-0.974)	AD1/AD6
[2]	0.926	(0.816-0.974)	AD2/AD5
[3]	0.914	(0.737-1.000)	A2/AD4
[4]	0.844	(0.553-0.947)	AD1/AD3/AD6
[5]	0.106	(0.026-0.395)	AD2/AD3/AD5

D homoeologs

	CF	CI	
[1]	0.981	(0.912-1.000)	AD2/AD5
[2]	0.778	(0.618-0.912)	AD1/AD6
[3]	0.727	(0.500-0.882)	D5/AD4
[4]	0.704	(0.559-0.794)	AD1/AD3/AD6
[5]	0.214	(0.147-0.324)	AD2/AD3/AD5
[6]	0.151	(0.059-0.382)	D5/AD2/AD5
[7]	0.133	(0.029-0.235)	D5/AD1/AD4
[8]	0.077	(0.029-0.206)	AD1/AD4/AD6
[9]	0.058	(0.029-0.176)	D5/AD3/AD4

Figure4

Table 1. Species and accessions included in the present analyses. PI numbers are given for germplasm accessible through the National Plant Germplasm System (<http://www.ars-grin.gov>), and citations are given for the other accessions. Accessions marked with an asterisk were used in both A and D homoeologous concordance analyses, whereas those marked with “A” or “D” were used only in that respective concordance analysis. All accessions were included for all other analyses; however, sequence quality filters may have removed specific accessions from individual gene analyses.

Genome	Species	Accession	
A2	<i>G. arboreum</i> L.	AKA84014 ^A	CottonGen germplasm resources (Yu et al. 2013)
D5	<i>G. raimondii</i> Ulbrich	JFW ^D	
AD1	<i>G. hirsutum</i> L.	Cascot L7	PI 607181
		Maxxa	PI 540885
		TX1009	PI 163012
		TX1182	PI 529978
		TX2090	PI 501497
		TX2094 ^A	PI 501501
		TX2095 ^D	PI 501502
		TX674	PI 158562
AD2	<i>G. barbadense</i> L.	BF19	
		DP340	(Bowman et al. 2006)
		GB0074 ^D	PI 528330
		GB0303	PI 528151
		GB0414 ^A	PI 528259
		K101	PI 528337
		Phy76	PI 617048
		PS7	PI 560140
AD3	<i>G. tomentosum</i> Nutt ex Seem	AD3-7*	
		Kahoolawe	
		(local lab accession)	
AD4	<i>G. mustelinum</i> Miers ex Watt	JALee	
		W400*	
AD5	<i>G. darwinii</i> Watt	PW7 ^A	
		WB1200 ^D	
AD6	<i>G. ekmanianum</i> Wittmack	Seijo3922*	(Grover et al. 2014)

Table 2. Genes included in the present analysis, as annotated in the *G. raimondii* genome (Paterson et al. 2012). All listed genes were used for Neighbor-Joining and Maximum Likelihood analyses. For the Bayesian concordance analysis, those genes marked with an asterisk denote those included in both A and D concordance analyses, whereas those marked as “A” or “D” denote those included in the concordance analysis only for the listed homoeologous genome.

Chr 1	Chr 5	Chr 9
Gorai.001G089400*	Gorai.005G097800*	Gorai.009G052100*
Gorai.001G134900*	Gorai.005G142000*	Gorai.009G062300*
Gorai.001G143200*	Gorai.005G229600*	Gorai.009G174600*
Gorai.001G222400*	Gorai.005G237400	Gorai.009G182300 ^A
Gorai.001G255600*		Gorai.009G200600 ^A
Gorai.001G266600*	Chr6	Gorai.009G244800*
Gorai.002G052600*	Gorai.006G078900*	Gorai.009G253400*
Gorai.002G064500	Gorai.006G172800	
	Gorai.006G177800	Chr 10
Chr 2		Gorai.010G110000*
Gorai.002G215000 ^D	Chr 7	Gorai.010G124800 ^A
Gorai.002G218500*	Gorai.007G023500*	
	Gorai.007G029600*	Chr 11
Chr 3	Gorai.007G056100	Gorai.011G037900*
Gorai.003G025900	Gorai.007G065000 ^A	Gorai.011G193600 ^D
	Gorai.007G099300*	
Chr 4	Gorai.007G218100*	Chr 12
Gorai.004G125200*	Gorai.007G267900 ^A	Gorai.012G022000
Gorai.004G134600*		Gorai.012G045800
Gorai.004G196800*	Chr 8	Gorai.012G152400
Gorai.004G210600*	Gorai.008G097100	
Gorai.004G212200*	Gorai.008G155400*	Chr 13
Gorai.004G231100*	Gorai.008G200200*	Gorai.013G106600 ^A
	Gorai.008G217500 ^D	Gorai.013G139200 ^A
	Gorai.008G276100