**Insights into the evolution of the New World diploid cottons (*Gossypium,* subgenus *Houzingenia*) based on genome sequencing**

Corrinne E Grover1, Mark A Arick II2, Adam Thrash2, Justin L Conover1, Walker C Pett1, Tracy Heath1, William S Sanders2,3,4,Daniel G Peterson2, Jodi Scheffler5, Brian Scheffler5, and Jonathan F Wendel1

1Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA 50011 USA; correspondence jfw@iastate.edu

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

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

4The Jackson Laboratory, CT, USA

5Jamie Whitten Delta States Research Center, USDA-ARS, Stoneville, MS 38776, USA

**Abstract:**

I (Jonathan) will write this once the ms is finished and I have something close to final.

**Keywords**: phylogenomics, molecular evolution, transposable elements, hybridization, introgression, rate variation

The American, diploid “D-genome” cottons (subgenus *Houzingenia*) comprise a monophyletic clade of cytogenetically and morphologically distinct species largely distributed from Southwest Mexico to Arizona, with additional disjunct species distributions in Peru and the Galapagos Islands [1–4]. Included in the 13-14 species presently recognized in subgenus *Houzingenia* [5,6] is a source of cytoplasmic male sterility in cotton, *G. harknessii* Brandegee, as well as the model diploid, D-genome progenitor to wild and domesticated allopolyploid (AD-genome) cotton, *G. raimondii* Ulbrich (reviewed in Wendel and Grover 2015). The close relationship of *Houzingenia* species to the agronomically important polyploid cottons has stimulated considerable interest in their diversity, distribution, and phylogenetic relationships. Accordingly, many of the species in the subgenus are taxonomically well-understood, although phylogenetic relationships remain incompletely resolved.

Early taxonomists divided subgenus *Houzingenia* into two sections and six subsections; these species alignments have, for the most part, been reiterated in subsequent phylogenetic studies [3,7–11], at least at the subsectional level. The alignment of subsections into their present taxonomic circumscriptions, however, does not appear to represent natural clades. Several molecular datasets have been used to evaluate these relationships, including chloroplast restriction sites [10]; simple sequence repeat (SSR) and expressed sequence tag (EST)-SSR markers [12,13]; random amplified polymorphic DNA (RAPD) markers [14]; internal transcribed sequences (ITS) [3]; and few single-copy nuclear genes [3]. Relationships among the six subsections, however, remain unclear, with different studies yielding alternative topologies [3,7,9,15]. Determining the closest living relative of the D-genome ancestor to the polyploid, however, has become firmly established. Early morphological and cytogenetic comparisons using intergenomic hybrids identified *G. raimondii* as the closest living relative to the D-genome ancestor of polyploid cotton species (reviewed in [16]). Subsequent analyses have support this observation [7–9,15,17–19], with few conflicts (see, however, [11]), as reviewed in Wendel and Grover (2015).

One consequence of these many molecular investigations has been the discovery of instances of putative hybridization among the D-genome cottons [20], and, in one remarkable case (i.e., *G. gossypioides*), between a *Houzingenia* species and another, geographically isolated subgenus from Africa (either A-, B-, E-, or, F-genome; [11,20,21]). Most remarkably, *G. gossypioides* appears to have been multiply introgressed, with an early nuclear introgression event followed by a much later hybridization to a member of the *G. raimondii* lineage, resulting in chloroplast, if not further (and cryptic), nuclear introgression [21]. Cytoplasmic introgression, and possibly cryptic nuclear introgression, is also present in some populations of *G. aridum*; i.e., the Mexican Colima populations of *G. aridum* possess a *G. davidsonii*- or *G. klotzschianum*-like cytoplasm [3].

Early attempts at understanding the evolution of the repetitive fraction of the genus support the inference of African introgression in *G. gossypioides* [22]; however, little else is understood with respect to the evolution of the non-genic fraction of *Houzingenia*. The D-genome cottons possess the smallest nuclear genomes in the genus, ranging only ~1.11 fold, from 841 Mb – 934 Mb. Notably, the distribution of genome sizes among the subsections suggests that the subgenus has experienced differential growth and/or reduction in genome size among species; however, the details of this sequence gain and loss have not been characterized. While the differences in genome size are not dramatic, the transposable element (TE) types that have accumulated in *G. raimondii* are different from those that have achieved higher copy numbers in the remainder of the genus [23–25]. Furthermore, research comparing the two sister genera to cotton (i.e., *Kokia* and *Gossypioides*; [25]) reveals that their equivalent genome sizes belies a more dynamic scenario of repetitive sequence gain and loss. A similar conclusion was reached for the two A-genome (subgenus *Gossypium*) species, whose small change in genome size (~1.05X) masks differences in TE accumulation [24,25].

Here we reexamine phylogenetic relationships and evolution in the cotton subgenus *Houzingenia* using next-gen (Illumina) sequencing data. We leverage newly generated genome and plastome sequences, the first for most of the included species, to address questions surrounding genome evolution in a monophyletic group of closely related species. We characterize both the pace and patterns of molecular evolution of genes and repetitive sequences, evaluate the amount of divergence outside of genes, and describe the history of indels and single-nucleotide polymorphisms (SNPs). Finally, we revisit the phylogeny of the D-genome clade, providing insight into relationships among species and with respect to sequence gain and loss among closely related species. Our results represent a phylogenomic characterization of molecular evolution for a closely related set of plant species and provide resources for comparative research and for the cotton community at large.

**Results**

**Genome assemblies and annotation**

Approximately 22-65X raw coverage libraries were sequenced for at least one representative of each D-genome species (Table 1), resulting in an average of 169.4 M reads per accession. Quality filters further reduced the number of reads per sample to an average of 136.9 M (range: 87.2 – 260.2 M), representing an average of 33X coverage per sample. All accessions (except *G. thurberi* accession 2) were assembled via ABySS using multiple kmer values (see methods) and the assembly with the greatest E-size [26] was selected to represent each species. These representative assemblies were improved with the reference-based scaffolder Chromosomer [27] using the closely related *G. raimondii* genome [28], producing assemblies that range in size from 585 to 775 Mbp (average 643 Mbp) and cover 67 – 85% of each genome (Table 2). These metrics are comparable to those generated by the subgenus *Houzingenia*-derivedreference genome [28].

Assemblies from all accessions were annotated, resulting in between 20,522 and 45,244 gene models per accession (min=26,492 for improved assemblies), similar to the number of primary transcripts published for *G. raimondii* [28]*.* BUSCO [29] analysis recovered over 80% of BUSCOs from nearly 80% of the improved assemblies, where a gene was considered present if more than 67% of the gene was recovered from that accession. This suggests a general completeness of the gene space, with an average of 87% complete BUSCOs recovered from each accession and less than 3.5% redundancy on average (Table 3).

Chloroplast reads were also recovered from the raw data, representing an average of 3% (range: 1.46 – 7.27%) of the filtered sequencing reads. These were used in reference-guided assemblies against the published *G. hirsutum* chloroplast genome [30]. The chloroplast genome alignment (excluding positions with ambiguity in any sequence) size was 158,996 bp, comparable to previously published cotton chloroplast genomes [31,32]. Chloroplast sequences were retained for phylogenetic analyses, and are available under Genbank accessions MH477706 through MH477724.

**Phylogenetic relationships among New World Cottons**

Phylogenetic relationships among *Houzingenia* species were revisited using a concatenation of 7,595 dispersed nuclear genes containing a minimum of one accession per species (see filtering criteria in methods). After removing any alignment position with >10% ambiguity, >20.3 million nucleotides derived from all 13 chromosomes remained for 28 *Houzingenia* accessions and for the outgroup, *G. longicalyx* (subgenus *Longiloba*). Maximum likelihood reconstruction of the phylogenetic relationships among species largely recapitulate established section and subsection relationships (Figure 1). As previously reported, while both sections of the subgenus, i.e., *Houzingenia* and *Erioxylum,* exhibit polyphyly, the individual subsections are either monophyletic or monotypic (Figure 1). Species relationships are largely congruent with the most recent phylogenetic inferences for the subgenus using nuclear genes [3], both of which differ from the subgenus SSR dendrogram [33] in the placement of several taxa, including *G. raimondii*, *G. davidsonii*, and *G. gossypioides*.

Notably, one of the two *G. aridum* accessions included (D4-12C from Colima, Mexico; PI 530897) is placed basal within the arborescent cottons of subsection *Erioxylum* and not sister to the *G. aridum* accession from Jalisco [3]. This observation recapitulates that of Alvarez et al (2006), which used AFLPs to evaluate 143 individuals from 50 populations of subsection *Erioxylum* species and the related subsection, *Integrifolia*, which was previously identified as a source of cytoplasmic introgression in Colima *G. aridum* accessions [34]. Indeed, phylogenetic analysis of the entire chloroplast for *Houzingenia* species (Figure 2) concurs with previous chloroplast restriction site analysis [10], which suggest that the Colima *G. aridum* accession (D4-12C) has an *Integrifolia* derived cytoplasm. It is interesting to note that diversity analyses of subsection *Erioxylum* using SSR markers [33,35,36] suggest that the circumscription of *G. aridum* may include previously undescribed species, a potential alternative hypothesis to introgression. SNP analyses of the two *G. aridum* accessions included here suggest that the Colima accession does retain evidence of nuclear introgression; this was determined using an ABBA-BABA test [37,38] with both accessions of *G. aridum* (H1 and H2); *G. davidsonii* as the source of introgression (H3); and *G. gossypioides* as the ancestral state (outgroup). This analysis confirms ancient admixture resulting in introgression from a *G. davidsonii*-like speciesinto *G. aridum* Colima (Z=-3.64, representing significant deviation from the mean). To further characterize the extent of nuclear introgression in *G. aridum* Colima, we compared the number of inferred introgressed SNPs (i.e., derived SNPs shared between *G. aridum* Colima and *G. davidsonii*) against the number of SNPs where *G. aridum* Jalisco (non-introgressed) shares a derived state with *G. davidsonii*. This tabulation (Table 4) recapitulates the results of the ABBA-BABA test (chi-square p-value = 0) in confirming nuclear introgression from subsection Integrifolia into *G. aridum* from Colima. When the data are partitioned by chromosome, about half of the chromosomes show an excess of derived SNPs compared to their counterpart in the non-introgressed *G. aridum* from Jalisco (Table 4), indicating that perhaps the genomic distribution of surviving introgressed regions has been uneven. Although the number of genes showing derived SNPs, and hence a residue of introgression, is not significantly different between the two *G. aridum* accessions, the Colima *G. aridum* does exhibit an excess of SNPs *in genes* (p = 0.0015). The latter is important in that these SNPs, while limited, both have high confidence in their orthology and recapitulate the broader genomic conclusions of there being ancient nuclear introgression in the Colima populations of *G. aridum*.

In addition to the evidence for introgression into Colima *G. aridum*, comparison between the nuclear and chloroplast phylogenies recapitulates previous observations of Austroamericana-derived introgression in section Selera, i.e. *G. gossypioides*. *Gossypium gossypioides* is unusual within *Houzingenia* as it has likely undergone two separate instances of introgression: (1) the more recent chloroplast introgression noted here and elsewhere [10,20,21], and (2) nuclear introgression, as evidenced by the presence of African cotton-like ITS [11] and repetitive DNA [39]. Clear evidence of chloroplast-nuclear conflict is seen in the analyses here, congruent with previous observations, which is resolved when the putatively introgressed accessions are removed (data not shown). Evidence for nuclear introgression is less clear (see below) and warrants additional analyses involving more *Gossypium* species, which is beyond the scope of the present paper.

**Divergence amounts and rates of molecular evolution**

Divergence times were calculated for the thirteen extant *Houzingenia* species (Figure 1) using the *chronos* function in the R package {ape} and using the median divergence time for the Malvaceae, as described in Grover et al (2018). Subgenus *Houzingenia* diverged an estimated 6.58 mya from the remaining cotton subgenera (here represented by *Longiloba*), a value within prior estimates [40]. The basal-most radiation is represented by the divergence of *G. gossypioides* from the rest of the subgenus, approximately 2.56 mya (Figure 1), although we note that there may be additional error in this estimation arising from cryptic nuclear introgression in *G. gossypioides*. For this reason, the time estimates for all nodes (including *G. gossypioides*) were based on the estimated divergence time range for the next most basal node, which separates section *Erioxylum* subsection *Erioxylum* from the remaining subgenus. Most species are inferred to have diverged relatively recently, within the last 0.5-2 my, with the notable exception of *G. davidsonii* and *G. klotzschianum*, here estimated to share an ancestor that is an order of magnitude more recent than previously suggested by allozyme and chloroplast restriction site analysis [41]. Their near-identical nature is reflected in both their estimated nuclear branch lengths (0.0003 substitutions per site versus 0.0018 to 0.0065 on other terminal branches) and their rates of substitution (0.0000 to 0.0048 dS and 0.0000 dN; Table 5). While this close relationship between *G. davidsonii* and *G. klotzschianum* has been reported previously [41], this is the first modern estimate of genome-wide divergence between these two species.

Genome-wide rates of molecular evolution among *Houzingenia* species were calculated for all species comparisons (Table 5). As expected, pairwise synonymous mutation rates (dS, average = 0.0213 substitutions/site) were approximately an order of magnitude greater than the nonsynonymous mutation rates (dN, average = 0.0026; Table 5). Synonymous mutation rates varied from 0.0000 between the two extant members of subsection Integrifolia, *G. davidsonii* and *G. klotzschianum*, to 0.0287 between *G. aridum* and the basal-most diverging member of *Houzingenia*, *G. gossypioides*. When considering divergence time between species, the dS range narrows to between 0 and 0.017 substitutions/site/million years (my) with 94% of the comparisons falling between dS/my=0.009-0.013. A single dS comparison, *G. davidsonii* and *G. klotzschianum*, was less than this range. No pattern was evident in the four values that exceeded this range. Similarly, dN varied from 0.000 between *G. davidsonii* and *G. klotzschianum* to 0.0033 between *G. lobatum* and *G. gossypioides*, again reflecting the basal nature of *G. gossypioides*. When standardized by time, the range narrows to dN=0-0.0018, with 90% between dN=0.0011-0.0015. Again, the Integrifolia species occupied the lowest dN value; however, notably, the dN value for *G. turneri* versus *G. harknessii* was similarly small (dN=0.0002). This stands in contrast to the dS value for the pair, which was comparably large at dS=0.0148 (Table 5).

**Transposable element characterization**

Similar to previous reports [28], repetitive DNAs contribute roughly half of the total genome sequence for all species in subgenus *Houzingenia*, from an average of 39.4% in *G. harknessii* to 46.9% in *G. armourianum*. Like most flowering plants, a vast majority of this sequence is due to the prevalence of class II *gypsy* elements, which comprise 29.2 - 34.3% of the total genome size for any *Houzingenia* species (Figure 3). Multi-dimensional TE profile visualization using both log transformed and percent-genome size standardized counts showed considerable overlap among species, and even among subsections (Figure 4). Multivariate t-distribution confidence ellipses (as implemented in ggplot2) are drawn for each subsection, all of which overlap with at least one other subsection. Even those subsections where insufficient sampling precludes the generation of a confidence ellipse (i.e., *Selera* and *Integrifolia*), the plotted data points are contained within the occupied space of another subsection (Figure 4). *Selera*, for example, is contained within the confidence ellipse for both all other subsections, as is *Integrifolia*. Likewise, few repetitive elements (14 elements at p<0.5, 13 *gypsy* and 1 undefined) differ significantly in copy number among *Houzingenia* species. This apparent overlap in repetitive element profiles is also suggested by the relative amounts of each transposable element category among subsections (Figure 3).

To compare the overlap among subsections, we performed a Procrustes ANOVA with complex linear models, as implemented in the R package [geomorph] [42]. For this analysis, we compared each subsection using all representatives of that subsection as indicators of variance. Few comparisons showed statistically significant differences, with the patterns of repetitive abundance differing only between *Austroamericana* and *Caducibracteata* and between *Integrifolia* and *Selera* (p<0.05).

The absolute amount of sequence attributable to each type of TE category is similar among *Houzingenia* species and is distinguishable from the African subgenera, primarily for *gypsy* elements (Figure 3). The total amount of *gypsy* elements predicted for the African species is far greater (average 878 Mb versus 277 MB, respectively), which is expected given previous analyses of cotton transposable elements [23,43,44]. The total amount of predicted MULE/MuDR-like elements, however, is greater for *Houzingenia* (average 4.4 Mb versus 1.6 Mb in the African subgenera) even despite the large difference in genome size, an observation not previously reported. These patterns persist even when comparing TEs as a function of genome size (Figure 6), with two additional observations. First, the large error bars for *gypsy* amount in *G. raimondii* become more pronounced. Inspection of the total amounts for this species suggest that there is a single accession (*G. raimondii* accession D5-6; Table 2) that has remarkably more *gypsy* elements than the remaining conspecifics. While approximately 30% of *gypsy* clusters in *G. raimondii* accession 6 are found in excess (relative to the other accessions), less than quarter of these contribute > 1 Mb additional sequence, indicating minor to modest relative proliferation in most cases. Interestingly, however, a single *gypsy* cluster (cluster 78) comprises 4.8 Mb additional sequence in *G. raimondii* accession 6 relative to the conspecific with the closest amount (12.6 Mb in *G. raimondii* accession 6 versus 7.8 Mb in accession 8). The average for this cluster, including *G. raimondii* accession 6, is only 5.2 Mb. These observations suggest that the *gypsy* element represented by cluster 78 has been recently active in the *G. raimondii* genome, achieving significant success in at least one lineage.

Previous research on *G. raimondii* (subsection *Austroamericana*) demonstrated a relative lack of lineage-specific amplification with concomitant removal of a prolific cotton *gypsy* element as a mechanism for genome downsizing in *G. raimondii* (J. S. Hawkins et al. 2009). Congruent with these results, most of the clusters recovered here are composed primarily of “older” reads (68.6 – 78.6% per accessions), i.e., reads more divergent than expected for recently active transposable elements. Ancestral state reconstruction of individual clusters, however, demonstrate both amplification and removal concomitant with the inferred changes in overall genome size (Figure 1; Supplementary Figure 2). Most clusters are “older”, with 39% of clusters comprised solely of “older” repeats and the remaining clusters most frequently showing recent amplification in one to few lineages (Supplementary Figure 3).

**Genome differentiation via insertions and deletions**

Small-scale insertions and deletions are a common form of sequence variation, with the potential to alter regulatory as well as coding regions [45–48]. While this is particularly true for large-scale, TE-associated indels (e.g., transposable element insertions), the formation of smaller indels can also vary among related species [49–51] . Accordingly, we evaluated the extent of indel evolution among *Houzingenia* species, using the *G. raimondii* genome as the reference state and polarized using *G. longicalyx* (subgenus *Longiloba*). Phylogenetic analysis of coded indels as multistate characters (see methods) recapitulates the nuclear phylogeny, suggesting that indel formation largely corresponds to species relationships. In total, small indels were present at 1,149,943 positions in at least one of the 13 *Houzingenia* species (relative to the outgroup *Longiloba*). Within *Houzingenia*, indels distinguish one or more species at 761,746 locations.The range in number of these distinguishing indels per chromosome varies by over 31,000 events, from 40,747 indels on chromosome 12 to 72,303 on chromosome 9, the longest and smallest chromosomes, respectively. Relative to the length of each chromosome, the gap narrows to between 779 indels/Mb on chromosome 5 to 1,174 indels/Mb on chromosome 8, a difference of 395 indels/Mb. Indels ranged in size from 1 – 270 nucleotides, with an average of 6.2 nt/indel. While the size of the largest indel detected varied among chromosomes, the average indel size per chromosome ranged narrowly from 5.7 – 6.7 nt/indel (Table 7).

Among accessions and chromosomes, the number of indels/Mb is relatively similar (98 – 260 indels/Mb on *G. raimondii* chromosome 1 and *G. gossypioides* chromosome 6, respectively; Supplementary Table 1), but statistically distinct (chi2 p<0.01). Deletions generally outweigh insertions for each chromosome/accession combination, both with respect to number (twofold) and length (2.5 to 5-fold; SupplementaryTable 1). This results in a net loss of between 278 and 555 kb per accession (*G. raimondii* and *G. trilobum*, respectively; average = 439 kb). Compared to the rate of nucleotide substitution, the rate of indel events is much lower and is approximately equivalent among species (from 16 – 18 nucleotide changes per indel event; Supplementary Table 1). The rate of indel formation among chromosomes and accessions varies slightly more than the overall rate, from 14 to 23 substitutions per indel. While no obvious patterns exist in this respect, the basal most radiation, *G. gossypioides*, consistently has more indels relative to SNPs, possibly as a consequence of its introgressed history [11,21]. While our understanding of the pattern and rate of indel formation among species would be increased through whole genome alignment of higher quality, *de novo* genome sequences rather than the resequenced genomes utilized here, our preliminary data suggest that differences in small indel evolution may not have a significant effect on this scale. These results do, however, support the idea that small deletions may be able to partially counteract genome size growth by TE amplification and small insertions.

**Genome differentiation via copy number evolution**

Recently, the extent of variation in gene content within and among plant species has been conceptualized in terms of the “pan-genome”, which refers to the suite of genes present within or among closely related species [52–60]. Here we begin to evaluate the scope of a *Houzingenia*-specific pan-genome by modeling genic copy number evolution. Homologous gene clusters generated via OrthoFinder were used as input in Count [61], which has been developed to conduct evolutionary analyses of homologous family sizes in a phylogenetic context, including inferring the rate of gene gain and loss for each phylogenetic branch. We found that the inferred rate of loss for a given lineage was consistently greater than the rate of gain (with the exception of *G. turneri*). Among lineage rate variability was observed for both inferred losses and gains; however, the magnitude of variability in the inferred rate of losses was far greater (0.05 – 0.41 losses per branch) than in gains (0.00 – 0.13 gains/branch). Standardizing these rates to account for variability in nucleotide substitution rates (as a proxy for time) reduces the difference in variability between the rate of loss (0.06 - 0.31) and gain (0.00 - 0.25).

Since these summarized rates of loss and gain could be influenced by the effects of a few orthogroups, we performed a random resampling of the data and plotted the distribution for losses and gains relative to the observed rate (Figure 7). Generally, with the exception of *G. turneri*, the inferred rate of loss greatly exceeded the resampled range, indicating the presence of highly influential orthogroups. The inverse, however, was observed in the resampled gain data, where the inferred rates typically were less than the resampled range. These results suggest that the rate of gene loss and gain in these lineages may be sensitive to changes in family size for a few orthogroups. A caveat, however, is that these inferences are based on orthogroup membership, which are clusters of closely related genes (i.e., gene families). In most cases, these orthogroups will have few members; however, in some cases, orthogroup membership will rise to many members in some species, such that there is an order of magnitude difference between species for those clusters. Therefore, while these results indicate patterns that may exist in copy number evolution among closely related species, further analyses involving synteny to determine strict orthology are required to fully understand the nuances of copy number evolution across time and among lineages.

**Discussion**

The New World diploid cottons comprise a monophyletic assemblage of primarily Mexican, D-genome species that are of interest because of their involvement in origin of the allopolyploid (AD-genome) cottons, which include the commercially important species *G. hirsutum* (upland cotton) and *G. barbadense* (Pima cotton) (reviewed in Wendel and Grover, 2015). In addition, previous work has indicated that several species in the group have complex evolutionary histories involving cryptic interspecific hybridization and introgression (all earlier citations). Here we employed whole genome resequencing data to generate an average of ~170 M reads for 22 accessions representing all species in the subgenus, producing ~33X genomic coverage per sample (genome sizes are approximately 880 Mb). We used these data to reassess the phylogenetic history of the subgenus and provide a temporal framework for its diversification; to explore further earlier hypotheses of interspecific introgression; and to create a comparative molecular evolutionary perspective on genomic diversification in a group of 13-14 relatively closely related species. Each of these topics is discussed in turn below.

**Phylogenetic and geographic history of subgenus *Houzingenia*.**

Earlier investigations of phylogenetic relationships within the subgenus were based on relatively shallow genomic surveys (e.g., several nuclear genes, or cpDNA restriction site analysis) or incomplete taxonomic sampling. Here we used 7,595 nuclear genes from throughout the genome as well as whole-chloroplast genome sequences to reevaluate phylogenetic relationships. Our analyses generally support the earlier inferences, recapitulating established sectional and subsection relationships (Figure 1), and further documenting that neither taxonomic section, i.e., *Houzingenia* and *Erioxylum,* is monophyletic. Among the most notable inferences enabled by the phylogenetic reconstruction based on the 7,595 nuclear genes are the following: (1) *Gossypium gossypioides* is phylogenetically basal within the clade, as suggested by earlier work using rather limited genomic sampling. This is a rare, highly localized species from Oaxaca, with an unusual genomic composition that appears to reflect accumulated reticulations with other species (this topic addressed below). (2) The Mexican complex of arborescent species (to 10 or more meters in height) continues to hold as a monophyletic assemblage, notwithstanding accessions of *G. aridum* from Colima (this also addressed below). (3) The best model of the D-genome donor to allopolyploid (AD-genome cottons), i.e., the geographically disjunct *G. raimondii* from Peru (reviewed in Wendel and Grover, 2015), is well-nested within the subgenus and is phylogenetically sister to the remarkably disjunct Baja California-Galapagos Islands species pair *G. davidsonii* and *G. klotzschianum*; these three species are sister to the Arizona-Sinoloan disjunct species pair *G. thurberi* and *G. trilobum*. (4) The three species from Baja California and adjacent islands, *G. harknessii, G. turneri* and *G. armourianum*, comprise a monophyletic group distinct from the fourth Baja California species *G. davidsonii*, with the first two of these three sister to each other.

The foregoing phylogenetic synopsis evokes a historical biogeography scenario of repeated long-distance dispersals in addition to possible vicariance events that generate geographical disjunctions. It is noteworthy that the aggregate geographical range of the complex extends from southern Arizona to Peru, but with a phylogenetic history that is inconsistent with a single directional radiation across the landscape from any single ancestral home.

Further perspective on this history is provided by the dating analysis, as shown in Figure 1. We estimate that the subgenus diverged from the remainder of the genus approximately 6.58 mya, consistent with a recent study based on 78 concatenated protein-coding exons from the chloroplast genome [40]. Among the most notable features of the chronogram (Figure 1) is that even though the subgenus appears to have originated about 6.6 mya, all surviving species trace to a much more recent origin about 2.5 mya, or in the Pleistocene. Thus, over 4 million years of evolutionary history of this group is lost, in that no surviving clade traces to the long branch between the D-genome and the remainder of the genus. In addition, nearly all of the biodiversity in the group is more recent in origin, within the last 0.5 – 2.0 million years, suggesting a period of both rapid diversification as well as geographic dispersal extending from Arizona (*G. thurberi*) to the Galapagos Islands (*G. klotzschianum*) and Peru (*G. raimondii*). This temporal framework emphasizes the remarkable and mysterious propensity for long-distance dispersal in the genus *Gossypium*, as reviewed elsewhere [6].

**Phylogenetic incongruence and ancient hybridization**.

In our earlier work [3,10,11,20,21,62], we highlighted how comparisons between phylogenetic inferences drawn from nuclear vs. chloroplast sequences are illuminating with respect to ancient hybridization events. Especially pertinent here is the evolutionary history of *G. aridum* and *G. gossypioides*. With respect to the former, we noted previously that populations of this wide-ranging species from the single Mexican state of Colima share a chloroplast genome with the Baja California – Galapagos Islands species pair *G. davidsonii* and *G. klotzschianum,* whereas populations from the remainder of the range have a chloroplast genome that is phylogenetically included in the rest of the arborescent clade (which includes *G. laxum, G. schwendimanii,* and *G. lobatum*) [10,62]. This same incongruence is recapitulated in the present study, with the added twist that in the reconstruction based on the nuclear genome, *G. aridum* from Colima appears in a basal position in the arborescent clade, and is thus biphyletic within this group. At present it is unclear whether this position reflects cryptic taxonomic diversity within the group (see discussion in [6]), or if instead *G. aridum* from Colima was “dragged” to its basal position by nuclear introgression from the *G. davidsonii* and *G. klotzschianum* lineage (with which it share cpDNA genomes). In this respect, we highlight the results from an AFLP survey [62] using a broad sampling of 24 populations of *G. aridum* (including 4 from Colima) as well as the other relevant species, in which it was concluded that the Colima populations are both genetically distinct and contain a comparatively high frequency of AFLP fragments that otherwise are diagnostic of the cpDNA donor clade. Given the biogeographic proximity of Colima to Baja California and hence *G. davidsonii*, we proposed a history, supported here by whole genome (nuclear and chloroplast) sequence data and our dating analysis (Figure 1), of migration of one or more seeds from Baja California to the Colima coast, perhaps during the Pleistocene followed by hybridization and geographically localized nuclear introgression.

In addition to the evidence for introgression into Colima *G. aridum*, comparison between the nuclear and chloroplast phylogenies (Figure 2) recapitulates the previously observed striking incongruence between the nuclear and cpDNA placement of *G. gossypioides*.As described in detail in a series of earlier publications [39][10,20,21], *G. gossypioides* is basal within the subgenus *Houzingenia* in nuclear gene trees yet exhibits apparent introgression of repetitive sequences from a different *Gossypium* lineage from Africa. Moreover, and equally extraordinary, this rare species is also confirmed (Figure 2) as sharing a relatively recent cpDNA ancestry with the equally rare Peruvian endemic *G. raimondii* (the only species with which it will form fertile F1 hybrids [63,64]). Thus, *G. gossypioides* likely has undergone two separate instances of introgression: (1) the more recent chloroplast introgression, convincingly shown here for entire chloroplast genomes, and (2) nuclear introgression, as evidenced by the presence of African cotton-like ITS and repetitive DNAs [7,11,21,65]. This complex genomic history exemplifies how even isolated lineages in different continents (in this case Central America, South America, and Africa) may be linked by a series of remarkable, highly improbable, long-distance dispersal and interspecific hybridization events.

A final comment concerning *G. gossypioides* is that we failed to detect the putative “African” nuclear genomic introgression that is clearly demonstrated by genomic slot blots [22]. Although we do not recapitulate the observation of introgression in using repeat clustering, our analysis does not preclude African-like repeats in the *G. gossypioides* genome; our results indicate only that this phenomenon is not evident in the present analysis. Analysis of individual clusters fails to reveal any clusters where *G. gossypioides* is significantly different in copy number from the rest of *Houzingenia*. BLAST analysis of the repeats reported by Zhao et al (1998) suggest the closest cluster is *gypsy* cluster CL31 (72% coverage of AF060607.1); however, this cluster is not enriched in *G. gossypioides* versus the rest of *Houzingenia* (data at https://github.com/IGBB/D\_Cottons\_USDA). This lack of enrichment is also reflected when the repetitive clones from Zhao et al (1998) are used to mask each *Houzingenia* genome; that is, neither repetitive clone masks a greater fraction of the *G. gossypioides* genome than any of the other assembled genomes. At present, we cannot explain the different results obtained from these studies, apart from suggesting that the different analytical methods select for different genomic regions or sequence types.

**Widespread nuclear and chloroplast genome incongruence**.

One of the principal phylogenetic observations of this study is that reconstructions based on nuclear and cpDNA genomes are highly incongruent in a number of respects (Figure 2). Part of the reason for this may be a history of documented (e.g., *G. aridum, G. gossypioides*) as well as unobserved interspecific introgression. In addition, [corrinne, this is where the obvious discussion goes about short interior nodes and support levels; we will need to address this, for sure, for our own understanding and also to get it by reviewers].

**Molecular evolutionary patterns, processes, and rates.**

A primary purpose of this study was to generate genome-wide estimates of molecularly evolutionary patterns, rates, and processes that generate genomic variation. At present there are few comparable investigations in plants for the time-scale and taxonomic diversity encompassed by the present study (everyone – help me find good examples to discuss!).

***Protein evolution****.* With respect to genic evolution, we report a relatively narrow range of interspecific non-synonymous substitution (dN), averaging 0.0014 per million years, with a synonymous rate about an order of magnitude higher (Table 5). Thus, protein evolution *per se* is inferred to be quite slow, only about 1% every 7 million years. This compares … help, folks, other comparable genera out there. We are unaware of comparable estimates for other plant genera, but we expect that life-history features such as generation time (long in *Gossypium*) will be highly correlated with rates of protein evolution, as they are with rate variation in general (JW: I have good citations for this comment).

***Transposable elements and the repetitive fraction****.* Similar to previous reports for *Gossypium* [28,66–71], about half of the genomic space in the species studied here is occupied by transposable elements or their still-similar decaying footprints. As with most flowering plants, a majority of this sequence is due to the prevalence of class II *gypsy* elements, which comprise about one third of each of the genomes studied here (Figure 3). Relatively few repetitive elements differ significantly in copy number among thespecies (Figure 4), indicating a relative genomic stasis in TE content during the last 6.5 million years, and specifically during the last 2.0 million years during which most of the modern lineages evolved. In contrast, *gypsy* elements have proliferated in the A-genome diploids (Figure 3) and elsewhere in the genus [23] following their divergence from the D-genome. We conclude that the TE fraction of the D-genome diploid cotton genomes has been relatively quiescent, especially when compared to other genomes such as those of many grasses, where the repetitive fraction has a far more rapid turnover (everyone, good recent citations here). One exception to this generalization is for *G. raimondii* accession 6, in which the *gypsy* element represented by cluster 78 appears to have recently proliferated (Figure 6). This was a surprising finding, given the exceptionally low levels of nucleotide diversity in this species (Wendel, unpublished) and the small geographic range it occupies in a couple of river valleys in coastal Peru.

An additional comment about relative occupancy of different TE classes among subgenera is that while the absolute amount of sequence attributable to *copia* elements is similar among subgenera *Houzingenia, Gossypium,* and *Longiloba* (37.4 – 41.3 Mb, average), this element type represents a larger portion of the genome in *Houzingenia* than in the two larger-genome African subgenera. This observation reflects either a lack of *both* *copia* element colonization and degradation since divergence of the three subgenera (i.e., stasis of *copia* elements), or convergence of absolute amounts, in a manner that conceals the dynamics of element turnover. Ancestral state reconstructions (images at https://github.com/IGBB/D\_Cottons\_USDA) suggest that the latter is more likely, as both reduction and increase in copy numbers for the annotated *copia* elements are observed, both for the *Houzingenia* species, as well as for the African species (represented by *Longiloba*). Within the D-genome diploids, dynamics of *copia* elements were characterized with respect to their aggregate effect on genome size (Supplementary Figure 1), using the reconstructed ancestral amount for each as a baseline; while *copia* elements comprise a higher proportion of the genome for *Houzingenia* species than for other cottons surveyed (Figure 5), these elements generally seem to be in decline (Table 6), as 65% of accessions experienced a net loss attributable to *copia* elements. This may be due in part to a paradox of TE proliferation; i.e., as an element achieves transpositional “success”, the number of homologous regions visible to the recombination-based deletional mechanisms also increases.

***Genome differentiation via insertions and deletions***. Small-scale insertions and deletions are a common form of sequence variation [49–51](everyone, bring to our attention great examples from plants, divided into within species vs. between species). To evaluate this form of molecular evolution in subgenus *Houzingenia,* we used the *G. raimondii* genome as the reference genome and polarized insertions vs. deletions using *G. longicalyx* (subgenus *Longiloba*) as a phylogenetic outgroup. We report 1,149,943 positions associated with an indel in at least one of the 13 *Houzingenia* species, and of these, indels distinguish one or more *Houzingenia* species at 761,746 locations.Although indels were found genome-wide, there was considerable variation by chromosome, ranging from 779 indels/Mb on chromosome 5 to 1,174 indels/Mb on chromosome 8, a difference of 395 indels/Mb. Most indels were small, averaging 6.2 nucleotides, with a range in size of 1 – 270 nucleotides (Table 7). It is likely that some larger indels were missed due to genome sequence incompleteness and because only one species was used as a reference genome.

One notable feature of these data is the observed bias toward deletions over insertions, which averages about twofold in number but 2.5-to-5-fold in length (SupplementaryTable 1). The net effect of these dynamics is genome downsizing, with an estimated net loss of about 0.44 Mb per species, with a range between 278 and 555 kb per accession. This observation supports the idea that small deletions may be able to partially counteract historical genome size expansion that originated from TE amplification. Because species in subgenus *Houzingenia* have the smallest genomes in the genus (in which diploids vary about threefold in genome size from ~850 - ~2700 Mb), these data suggest that the process of genomic pruning remains active today, or at least it has been in the recent past. Finally, our comparative genomic data reveal, at finest scale of aligned nucleotides, a dynamic process of genomic downsizing that we inferred from computational modeling a decade ago [43]. Everyone: examples from other plants, for comparative context?

**Conclusions**

Not sure what I might write here, but perhaps a synopsis that is different from the abstract. Better, though, would be a comment that is forward-looking, about how the kind of analyses presented here will soon be commonplace, at which point we will not only gain much deeper phylogenetic insights, but also new clarity about modes and mechanisms and rates of molecular/genomic evolution. Take a stab at this or part of this if you feel the urge or inspiration.

**Materials and Methods**

*Sequence generation and initial processing*. DNA was extracted from (LEAVES) using (WHAT KIT), and sent to (WHERE) for library construction and sequencing. Sequencing was completed on the Illumina (WHAT MACHINE) using (WHICH SEQUENCING). Reads were trimmed and filtered with Trimmomatic v0.32 [72] 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. Detailed parameters can be found at https://github.com/IGBB/D\_Cottons\_USDA.

*Genome assembly and annotation*. Trimmed data were independently assembled for each species via ABySS v2.0.1 [73], using every 5th kmer value from 40 through 100. A single assembly with the highest E-size [26] was selected for each species and subsequently annotated with MAKER v2.31.6 [74] using evidence from: (1) the NCBI *G. raimondii* EST database [citation], (2) *G. raimondii* reference genome predicted proteins, as hosted by CottonGen.org [28], and (3) three *ab initio* gene prediction programs, i.e. Genemark v4.30 [75], SNAP v2013-11-29 [76], and Augustus v3.0.3 [77]. Both the SNAP and Augustus models were trained using BUSCO v2.0 [29]. Chromosomer version 0.1.3 [27], a reference-assisted scaffolder, was used to scaffold the selected assemblies against the gold standard *G. raimondii* genome. MAKER v2.31.6 [74] was used to transfer the previous annotations to the Chromosomer-based scaffolds by rerunning MAKER and using the transcripts from the original annotation as evidence.

**Phylogenetic analyses and ancestral state reconstruction.**

Trimmed reads from the genome assembly were mapped against the *G. raimondii* reference sequence [28] using BWA v0.7.10 [78], post-processed with samtools [79], and individual genes were independently assembled for each species/accession via BamBam v 1.3 [80] in conjunction with the *G. raimondii* reference annotation [28]. Alignments were pruned for genes and/or alignment positions with insufficient coverage, i.e., too many ambiguous bases, using filter\_alignments (https://github.com/IGBB/D\_Cottons\_USDA/). Parameters were set to remove sequences with more than 10% ambiguous bases within species and to remove aligned positions with more than 10% ambiguity among species. Genes were additionally filtered by length, to retain only those genes between a minimum of 500 bp and a maximum of 4051 bp, the latter of which represents the *G. raimondii* genome-wide mean plus three standard deviations. Only those genes with a minimum of one accession per species were retained for phylogenetic and molecular analyses. Genes were concatenated and subjected to maximum likelihood (ML) analysis via RaxML [81] using the basic general time reversible model with gamma distribution (GTRGAMMA), 10000 alternative runs on distinct starting trees, and rapid bootstrapping with consensus tree generation. The ML trees were rooted with a member of subgenus *Longiloba*, *G. longicalyx* (African F-genome).

Molecular evolutionary analyses were conducted in R v3.4.4 [82]. Species divergence time estimates were calculated via chronoMPL from {ape} [83], using the divergence estimates previously calculated for the Malvaceae [25]. Trees were visualized using the {ape} package [83]. Ancestral state reconstructions for genome size were completed using fastAnc from {phytools} [84]. Indels and SNPs were characterized among Houzingenia using the Genome Analysis ToolKit [85] and the *G. raimondii* reference sequence [28]. SNP introgression was measured both by individual SNP proportions (see https://github.com/IGBB/D\_Cottons\_USDA/) and ANGSD [37].

Gene orthology among species was determined via OrthoFinder [86], and copy numbers per species/gene family was derived from the resulting orthologous clusters. Copy number evolution was modeled using Count [61], and bootstrap-like replicates were modeled by resampling (with replacement) for 1000 permutations.

**Repetitive sequence characterization**

Reads from only one of the paired-end files (i.e., R1) were filtered and trimmed via Trimmomatic version 0.33 [87] to a uniform 85 nt (https://github.com/IGBB/D\_Cottons\_USDA), and then randomly subsampled to represent a 1% genome size equivalent (GSE) for each individual [88,89]. These 1% GSEs were combined as input into the RepeatExplorer pipeline [90,91], which has been successfully used to profile genomic repeats using low-coverage, short read sequencing. Only clusters which contain at least 0.01% of the total input sequences (i.e., 387 reads from a total input of 3,872,016 reads) were retained for annotation as per Grover 2018 [25], which uses the RepeatExplorer implementation of RepeatMasker [92] and a custom cotton-enriched repeat library. Genome occupation of each broad repeat type was calculated (in megabases; Mb) for each genome/accession based on the 1% genome representation of the sample and the standardized read length of 85 nt.

Patterns of repeat content per genome were determined using the abundance of each cluster in a multivariate dataset. Initial visualization of the data was conducted in R [82] using Principle Coordinate Analysis on read counts, either log normalized (to compare overall patterns of repeats) or normalized by genome size (to compare proportional cluster size). Differential abundance in cluster occupation was iteratively calculated at increasing phylogenetic depths to understand the evolution of repeat types at different temporal scales; that is, differentially abundant clusters were determined (1) within species, (2) between sister taxa, and (3) between deeper phylogenetic nodes. For each cluster, the ancestral state was reconstructed and used for comparison in the next analysis. Ancestral state reconstructions were completed using fastAnc for reconstruction [84] and the fitContinuous function of {Geiger} [93] for visualization. All analyses are available at (https://github.com/IGBB/D\_Cottons\_USDA).

*Repeat heterogeneity and relative age*. Relative cluster age was approximated using the among-read divergence profile of each cluster, as previously used for *Fritillaria* [94], dandelion [95], and *Kokia*/*Gossypioides* [25], sister outgroup genera to *Gossypium*. Briefly, cluster-by-cluster all-versus-all BLASTn [96,97] searches were conducted using the same BLAST parameters implemented in RepeatExplorer. A pairwise percent identity histogram was generated for each cluster, and regression models were used to describe the trend (i.e., biased toward high-identity, “young” or lower-identity, “older” element reads) using Bayesian Information Criterion [98] to select the model with the most confidence; specific parameters can be found in [25] and at https://github.com/IGBB/D\_Cottons\_USDA. The read similarity profile was automatically evaluated for each cluster to determine if the reads trend toward highly similar “young” or more divergent “older” reads. These profiles generally consist of six different trends: (1) positive linear regression ("young"); (2) absence of linear regression ("old"); (3) negative linear regression ("old"); (4) positive quadratic vertical parabola, trend described by right-side of vertex ("young"); (4b) positive quadratic vertical parabola, trend described by left-side of vertex ("old"); (5) negative quadratic vertical parabola, trend described by right-side of vertex ("old"); and (6) negative quadratic vertical parabola, trend described by left-side of vertex and vertex at >99% pairwise-identity ("old"). We note that young" and "old" are relative designations and not indicative of absolute age.

**Acknowledgements**Li Wang and Matthew Hufford

ResearchIT

**Figure legends**

Figure 1: Nuclear phylogeny of *Houzingenia* without (left) and including (right) the introgressed accession of *G. aridum* from the Mexican state of Colima. Divergence times are visualized on an ultrametric tree (left) whose colors correspond to the relative growth (blue) or reduction (red) of genome size in *Houzingenia*, as compared to the outgroup *G. longicalyx* (*Longiloba*). Inferred ancestral genome sizes are displayed on a proportional tree (right) whose colors correspond to the degree of change within *Houzingenia* alone. Phylogenetic methods, divergence time estimates, and ancestral state reconstruction details are found in the methods.

Figure 2: Comparison of phylogeny from reference-guided assembly of chloroplast-derived reads in *Houzingenia* (left; ML-derived branch lengths are listed) and the nuclear phylogeny (right). The chloroplast phylogeny shown here was derived from WGS-derived whole chloroplast sequences (see methods); however, this tree topology is recapitulated in a concatenated chloroplast gene-only phylogenetic reconstruction that includes all published sequences in Genbank (see <https://github.com/IGBB/D_Cottons_USDA> for details). Each node in the chloroplast phylogeny had 100% bootstrap support. While within subsection associations among species are recapitulated between the trees (in colors), the relationship among subsections varies between the two molecule types.

Figure 3: Mean transposable element content for each category in each species of *Houzingenia*, as well as representatives from *Gossypium* and *Longiloba*. The (average) aggregate number of kilobases represented by each transposable element category for each species (genome sizes included next to species names). Transposable elements were broadly categorized into categories and their representation per species summarized, with the minimum and maximum per species included.

Figure 4: Multidimensional comparison of *Houzingenia* species based on repetitive content. Species are designated by their numbered designations; D1 (*G. thurberi*), D2-1 (*G. armourianum*), D2-2 (*G. harknessii*), D3D (*G. davidsonii*), D3K (*G. klotzschianum*), D4 (*G. aridum*), D5 (*G. raimondii*), D6 (*G. gossypioides*), D7 (*G. lobatum*), D8 (*G. trilobum*), D9 (*G. laxum*), D10 (*G. turneri*), and D11 (*G. schwendimanii*).

Figure 5: Multidimensional comparison of *Houzingenia* species with African representatives from *Gossypium* and *Longiloba*. Species are again designated by their numbered designations; D1 (*G. thurberi*), D2-1 (*G. armourianum*), D2-2 (*G. harknessii*), D3D (*G. davidsonii*), D3K (*G. klotzschianum*), D4 (*G. aridum*), D5 (*G. raimondii*), D6 (*G. gossypioides*), D7 (*G. lobatum*), D8 (*G. trilobum*), D9 (*G. laxum*), D10 (*G. turneri*), D11 (*G. schwendimanii*), A1 (*G. herbaceum*), A2 (*G. arboreum*), and *F1* (*G. longicalyx*).

Figure 6: Mean transposable element content for each category in each species of *Houzingenia*, as well as representatives from *Gossypium* and *Longiloba*, relative to their genome size. Transposable elements were broadly categorized into categories and their representation per species summarized, with the minimum and maximum per species included.

Figure 7: Rate of gene gain or loss, per million years. Boxplot distributions show distribution of gene gain (A) or loss (B), per species, as inferred from the resampled data (see methods). Inferred rates of gain or loss from the total dataset are displayed as green triangles. Inferred rates for both gain and loss are substantially higher in *G. davidsonii* and *G. klotzschianum*, likely due to rate inflation based on the substantially shorter branches leading to these taxa.

Supplementary Figure 1: Magnitude of overall sequence gain and loss attributable to *copia* elements, per accession. Inset into the figure is the distribution of the number of accessions and clusters that have either gained (right) or lost (left) *copia* sequence.

Supplementary Figure 2: Ancestral state reconstruction of the fourteen clusters that distinguish at least one *Houzingenia* species from the others. Clusters represented are (left to right): CL0030, CL0053, CL0074, CL0086, CL0142, CL0154, CL0167, CL0193, CL0201, CL0301, CL0339, CL0341, CL0344, and CL0357. All clusters are gypsy elements, save for CL0301, which was unidentifiable.

Supplementary Figure 3: Cluster age distribution in *Houzingenia*. (A) The total number of clusters categorized as “old” or “young” in each *Houzingenia* species. The number of “old” clusters typically outweighs the number of “young” by two- to three-fold. (B) The number of clusters (y-axis) with a given number of accessions categorized as young (x-axis).