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

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

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

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

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

4 The Jackson Laboratory, CT, USA

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

Abstract

Keywords

**Background**

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 (Endrizzi, Turcotte, and Kohel 1985)added marginal citations. Included in the 13-14 species currently included in subgenus *Houzingenia* (Ulloa et al. 2013; Wendel and Grover, 2015) 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 (R C Cronn et al. 1996; Seelanan, Schnabel, and Wendel 1997; R L Small and Wendel 2000; J F Wendel and Albert 1992; J F Wendel, Schnabel, and Seelanan 1995; Álvarez, Cronn, and Wendel 2005), 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 [Wendel and Albert, 1992]; simple sequence repeat (SSR) and expressed sequence tag (EST)-SSR markers [citation]; random amplified polymorphic DNA (RAPD) markers [citation]; internal transcribed sequences (ITS) [Alvarez et al., 2005]; and few single-copy nuclear genes [Alvarez et al., 2005]. Relationships among the six subsections, however, remain unclear, with different studies yielding alternative topologies (R C Cronn et al. 1996; Liu et al. 2001; R L Small and Wendel 2000; Alvarez et al., 2005). 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 J F Wendel and Cronn 2003). Subsequent analyses have support this observation (Abdalla et al. 2001; Richard C Cronn, Small, and Wendel 1999; Liu et al. 2001; R C Cronn et al. 1996; Seelanan, Schnabel, and Wendel 1997; Randall L Small et al. 1998; R L Small and Wendel 2000), with few conflicts (see, however, (J F Wendel, Schnabel, and Seelanan 1995)), 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 (R. Cronn and Wendel 2004), 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; (Wendel et al., 1995; R. Cronn and Wendel 2004; R. Cronn et al. 2003)). 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 (R. Cronn et al. 2003). 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 (Alvarez et al citation here).

Early attempts at understanding the evolution of the repetitive fraction of the genus support the inference of African introgression in *G. gossypioides* (Zhao et al. 1998); 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 than those that have achieved higher copy numbers in the remainder of the genus (Jennifer S Hawkins et al. 2006; Corrinne E Grover et al., n.d.; Renny-Byfield et al. 2016; anything in the genome sequencing papers here?). Furthermore, research comparing the two sister genera to cotton (i.e., *Kokia* and *Gossypioides*; (Corrinne E Grover et al., n.d.)) 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 (Renny-Byfield et al. 2016; Corrinne E Grover et al., n.d.).

Here we reexamine phylogenetic relationships and evolution 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 addressing 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 also 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 (Salzberg et al. 2012) was selected to represent each species. These representative assemblies were improved with the reference-based scaffolder Chromosomer (Tamazian et al. 2016) using the closely related *G. raimondii* genome (Paterson et al. 2012), 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 (Paterson et al. 2012).

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* (Paterson et al. 2012)*.* BUSCO (Simão et al. 2015) 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 (Lee et al. 2006). The chloroplast genome alignment (excluding positions with ambiguity in any sequence) size was 158,996 bp, comparable to previously published cotton chloroplast genomes (Chen et al. 2016, 2017; Richard C Cronn et al. 2002). Chloroplast sequences were retained for phylogenetic analyses, and are available under WHAT-PRJNA

**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 (Álvarez, Cronn, and Wendel 2005), both of which differ from the subgenus SSR dendrogram (Ulloa 2014) 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 (Álvarez, Cronn, and Wendel 2005). 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 (Dejoode 1992). Indeed, phylogenetic analysis of the entire chloroplast for *Houzingenia* species (Figure 2) concurs with previous chloroplast restriction site analysis (J F Wendel and Albert 1992), 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 (Ulloa 2006, Ulloa 2014, Feng 2011) 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. To evaluate the presence of nuclear introgression, we conducted an ABBA-BABA test (Korneliussen, Albrechtsen, and Nielsen 2014; Sousa and Hey 2013) using 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 the 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*. *G. 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 (Wendel et al., 1992; R. Cronn et al. 2003; R. Cronn and Wendel 2004), and (2) nuclear introgression, as evidenced by the presence of African cotton-like ITS (J F Wendel, Schnabel, and Seelanan 1995) and repetitive DNA (Zhao et al. 1998). 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 per Grover et al (2018); see methods. Subgenus *Houzingenia* diverged an estimated 6.58 mya from the remaining cotton subgenera (here represented by *Longiloba*), which falls within prior estimates for the basal radiation of cotton lineages (Senchina et al. 2003). The basal-most radiation of the genus 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 (J F Wendel and Percival 1990). 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 (J F Wendel and Percival 1990), 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 (Paterson et al., 2012, others??), repetitive DNAs contribute roughly half of the total genome sequence for all species in subgenus *Houzingenia*, from an average of 44.5% in *G. lobatum* to 52% 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 32.7 - 37.9% 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 *Caducibracteata* and *Houzingenia*; likewise, *Integrifolia* is within *Houzingenia* and *Austroamericana*. Likewise, few repetitive elements (15 elements at p<0.5, 13 *gypsy* and 2 *copia*) 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.

To formally compare the overlap among subsections, we performed a Procrustes ANOVA with complex linear models, as implemented in the R package [geomorph]. 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 both *Caducibracteata* and *Erioxylum* (p<0.05). Interestingly, the variation in repetitive elements found in monotypic *Selera*, i.e., *G. gossypioides*, was not distinct from the remainder of subgenus *Houzingenia*. This stands in contrast to previous reports (Zhao et al. 1998), which noted the presence of repeats derived from "African cottons" (here represented by subgenera *Gossypium* and *Longiloba*, i.e., A- and F-genome species). This result is further apparent when including the African subgenera in the ordination (Figure 5); that is, *G. gossypioides* is clearly lumped with the other *Houzingenia* species. While this results stands in contrast to expectations, given previous reports (citations again here?), this analysis does not preclude African-like repeats in the Oaxacan endemic *G. gossypioides* genome; our results indicate only that this phenomenon is not evident in the present analysis. Analysis of individual clusters reveals only two clusters (i.e., *gypsy* clusters 186 and 368) where *G. gossypioides* is significantly different in copy number from the rest of *Houzingenia*, and *G. gossypioides* is also significantly different from the African representatives for these clusters. Copy numbers for those clusters suggest an increase in copy number for *G. gossypioides* relative to both the rest of *Houzingenia* and the African subgenera, and therefore are not part of the repeats that distinguish *G. gossypioides* from the African subgenera, as reported by Zhao et al (1998) and in the ribosomal sequences as reported by Wendel et al. (1995). BLAST analysis of the repeats reported by Zhao et al (1998) suggest the closest cluster is *gypsy* cluster CL63 (e-value: 4e-100); however, this cluster is not enriched in *G. gossypioides* versus the rest of *Houzingenia* (average 161.5 kb versus average 214.7 kb, respectively). 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.

The absolute amount of sequence attributable to each type of TE category is similar among *Houzingenia* species and is distinguishable from the African subgenera only in the amount of putative *gypsy* and MULE/MuDR-like elements (Figure 3) in opposing directions. The total amount of *gypsy* elements predicted for the African species is far greater (average 907 Mb versus 309 MB, respectively), which is expected given previous analyses of cotton transposable elements (Jennifer S Hawkins et al. 2006; J. S. Hawkins et al. 2009; C.E. Grover et al. 2007). The total amount of predicted MULE/MuDR-like elements, however, is greater for *Houzingenia* (average 6.3 Mb versus 2.4 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 35% of *gypsy* clusters in *G. raimondii* accession 6 are found in excess (relative to the other accessions), less than half contribute > 1 Mb additional sequence. Interestingly, however, a single *gypsy* cluster (cluster 62) comprises 11.3 Mb additional sequence in *G. raimondii* accession 6 relative to the conspecific with the closest amount (19.7 Mb in *G. raimondii* accession 6 versus 8.4 Mb in accession 8). The average for this cluster, including *G. raimondii* accession 6, is only 6.7 Mb. These observations suggest that the *gypsy* element represented by cluster 62 has been recently active in the *G. raimondii* genome, achieving significant success in at least one lineage.

A second notable observation from the genome-size standardized TE amounts is that while the amount of sequence attributable to *copia* elements is similar among subgenera *Houzingenia, Gossypium,* and *Longiloba*, this element type represents a larger portion of the genome in *Houzingenia* than in the two 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 an actual dynamic of element turnover. Ancestral state reconstructions (see images at GITHUB repo) 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*). Dynamics of *copia* elements in *Houzingenia* were broadly characterized with respect to their aggregate effect on genome size (Supplementary Figure 1), using the reconstructed ancestral amount for each as a baseline. Interestingly, 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 67% of accessions experienced a net loss attributable to *copia* elements.

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 (78%; range 76.5 – 79.7% per species) are composed primarily of “older” reads, 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). Generally speaking, however, most clusters are “older” and even the most recently amplified repeats are inferred to be active at the base of the subgenus (Supp Figure all histogram), with a small number (20) that appear to be recently active in terminal lineages.

**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 (Britten et al. 2003). 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 citations. 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 214,222 positions in at least one of the 24 accessions sequenced. The range in number of indels per chromosome varies by over 7,800 events, from 11,703 indels on chromosome 12 to 19,586 on chromosome 9; however, relative to the length of each chromosome, the gap narrows to between 227 indels/Mb on chromosome 5 to 330 indels/Mb on chromosome 12. Still, this difference in indel number per megabase is striking and significant (statistical test, help!).

Generally speaking, insertions outweigh deletions for each chromosome/accession combination, with the exception of chromosomes 8 and 9 for subsection Integrifolia (*G. klotzschianum* and *G. davidsonii*; Table Indels). Chromosome 9, in particular, appears susceptible to deletions, as the only chromosome where the total length in deletions outweigh insertions for more than half (60%) of the accessions. While the total amount of sequence change is relatively small (maximum gain = 61 kb and maximum loss = -31 kb), this belies the number (average 106,016 events per accession) and amount of sequence involved (average 540,196 inserted or deleted).

Compared to the rate of nucleotide substitution, the rate of indel events is much lower and is approximately equivalent among species (from 5.1 – 6.4 nucleotide changes per indel event), with the exception of the *G. raimondii* accession; however, as these indels and substitution rates are relative to the *G. raimondii* reference genome, SOMETHING. The rate of indel formation among chromosomes and accessions varies slightly more than the overall rate, from 3.79 to 8.52 substitutions per indel. While no obvious patterns exist in this respect, the two species from subsection *Integrifolia* tend to have more indels per nucleotide substitution whereas *G. gossypioides* tends to have fewer indels per nucleotide substitution. Although distance from the reference is considered here, it is notable that these patterns also coincide with the distance of these species from the reference. 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 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.

**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 (multiple refs). Here we begin to evaluate the scope of a *Houzingenia*-specific pan-genome by modeling copy number evolution. Homologous gene clusters generated via OrthoFinder were used as input in Count (Csurös 2010), 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 here that the inferred rate of loss for a given lineage was consistently greater than the rate of gain (with the exception of *G. turneri*), as might be expected. 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 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 loss/gain boxplots). 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**

**Evolution of coding regions: content variation and molecular evolution among closely related species**

**Evolution of noncoding regions: SUBHEADING (includes TEs and indels)**

D6 repeats – why didn’t we find Af specific repeats? Idk. Only 2 A-genome like repeats were reported in D6 (pxp 095 and pxp271). Of Andy’s repeats, first is A-specific and the second is A-enriched. Neither hits any proteins by blastx

**Conclusions**

**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 [citation] 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. [Let's port this repo to a lab site after and give the new url]

*Genome assembly and annotation*. Trimmed data were independently assembled for each species via ABySS v2.0.1 (Simpson et al. 2009), using every 5th kmer value from 40 through 100. A single assembly with the highest E-size (Salzberg et al. 2012) was selected for each species and subsequently annotated with MAKER v2.31.6 (Holt and Yandell 2011) using evidence from: (1) the NCBI *G. raimondii* EST database [citation], (2) *G. raimondii* reference genome predicted proteins, as hosted by CottonGen.org (Paterson et al. 2012), and (3) three *ab initio* gene prediction programs, i.e. Genemark v4.30 (Genes et al. 2003), SNAP v2013-11-29 (Korf 2004), and Augustus v3.0.3 (Stanke et al. 2006). Both the SNAP and Augustus models were trained using BUSCO v2.0 (Simão et al. 2015). Chromosomer version 0.1.3 (Tamazian et al. 2016 10.1186/s13742-016-0141-6), a reference-assisted scaffolder, was used to scaffold the selected assemblies against the gold standard *G. raimondii* genome and transfer the previous annotation to the new scaffolds.

**Gene stuff**

Gene orthology and family designations were determined via OrthoFinder [citation]...

**Phylogenetic analyses and ancestral state reconstruction.**

Trimmed reads from the genome assembly were mapped against the *G. raimondii* reference sequence (Paterson et al. 2012) using BWA v0.7.10 (Li and Durbin 2009), post-processed with samtools (which version) (Li et al. 2009), and individual genes were independently assembled for each species/accession via BamBam v 1.4 (Page, Gingle, and Udall 2013) in conjunction with the *G. raimondii* reference annotation (Paterson et al. 2012). 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 (cite Mirarab 2016) 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.

Species trees were estimated from individual gene trees via SNaQ (citation) and MP-EST (citation) using Bayesian and Maximum Likelihood analyses, respectively. Bayesian analyses were generated using MrBayes v (which version) (Huelsenbeck and Ronquist 2003) under GTR gamma with the following parameters: four runs with four chains for 1 million generations and using a burn-in fraction of 25%. Concordance among individual gene trees was assessed via BUCKy (Ané et al., 2007; Larget et al., 2010) with 3 runs, each with 4 chains and 1 million iterations, and default parameters. Quartet MaxCut was used to estimate the starting tree, and SNaQ was run using this starting tree and the concordance factors estimated by BUCKy. Visualization of networks….

Maximum likelihood (ML) analyses were performed using RaxML v(which version) (Stamatakis 2014) 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). MP-EST (citation) was used to estimate the species tree from the population of gene trees. Visualized how..

Molecular evolutionary analyses were conducted in R v(which version) (R Core Team 2017). Species divergence estimates were calculated via the [chronoMPL] package (citation), using the (which time estimates?) (citation). Trees were visualized using the [ape] package (Paradis, Claude, and Strimmer 2004). Ancestral state reconstructions for genome size were completed using [fastAnc]. Indels and SNPs were characterized among Houzingenia] using the Genome Analysis ToolKit (gatk) and the *G. raimondii* reference sequence (Paterson citation). Distance measures of aligned intergenic regions were estimated via [ape], and indels were characterized by (???).

**Repetitive sequence characterization**

Reads from only one of the paired-end files (i.e., R1) were filtered and trimmed via Trimmomatic version 0.33 (Bolger, Lohse, and Usadel 2014) to a uniform 95nt (https://github.com/IGBB/D\_Cottons\_USDA), and then randomly subsampled to represent a 1% genome size equivalent (GSE) for each individual (Hendrix and Stewart 2005; Jonathan F Wendel et al. 2002). These 1% GSEs were combined as input into the RepeatExplorer pipeline (Novák, Neumann, and Macas 2010; Novák et al. 2013), 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 (in revision), which uses the RepeatExplorer implementation of RepeatMasker (Smit, Hubley, and Green, n.d.) 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 95 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 using both calculated in R (R Core Team 2017) 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 (Revell 2012) and the fitContinuous function of [Geiger] (Harmon et al. 2008) 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 (Kelly et al. 2015), dandelion (Ferreira de Carvalho et al. 2016), and *Kokia*/*Gossypioides* (Corrinne E Grover et al., n.d.), sister outgroup genera to *Gossypium*. Briefly, cluster-by-cluster all-versus-all BLASTn (Boratyn et al. 2013; Camacho et al. 2009) 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 (Schwarz 1978) to select the model with the most confidence; specific parameters can be found in (Corrinne E Grover et al., n.d.) 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"; Figure WHATEVER). 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\_phylogeny: G. thurberi (By Katja Schulz from Washington, D. C., USA (Arizona Wild Cotton) [CC BY 2.0 (http://creativecommons.org/licenses/by/2.0)], via Wikimedia Commons) – also used = D10, D4, D3d