11-23-2018

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

Corrinne E. Grover  
*Iowa State University*, corrinne@iastate.edu

Mark A. Arick II  
*Mississippi State University*

Adam Thrash  
*Mississippi State University*

Justin L. Conover  
*Iowa State University*, jconover@iastate.edu

William S. Sanders  
*Mississippi State University*

Follow this and additional works at: [https://lib.dr.iastate.edu/eeob_ag_pubs](https://lib.dr.iastate.edu/eeob_ag_pubs)

Part of the *Ecology and Evolutionary Biology Commons*, *Molecular Genetics Commons*, and the *Plant Breeding and Genetics Commons*

The complete bibliographic information for this item can be found at [https://lib.dr.iastate.edu/eeob_ag_pubs/321](https://lib.dr.iastate.edu/eeob_ag_pubs/321). For information on how to cite this item, please visit [http://lib.dr.iastate.edu/howtocite.html](http://lib.dr.iastate.edu/howtocite.html).

This Article is brought to you for free and open access by the Ecology, Evolution and Organismal Biology at Iowa State University Digital Repository. It has been accepted for inclusion in Ecology, Evolution and Organismal Biology Publications by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Insights into the evolution of the New World diploid cottons (Gossypium, subgenus Houzingenia) based on genome sequencing

Abstract
We employed phylogenomic methods to study molecular evolutionary processes and phylogeny in the geographically widely dispersed New World diploid cottons (Gossypium, subg. Houzingenia). Whole genome resequencing data (average of 33X genomic coverage) were generated to reassess the phylogenetic history of the subgenus and provide a temporal framework for its diversification. Phylogenetic analyses indicate that the subgenus likely originated following trans-oceanic dispersal from Africa about 6.6 mya, but that nearly all of the biodiversity evolved following rapid diversification in the mid-Pleistocene (0.5-2.0 mya), with multiple long-distance dispersals required to account for range expansion to Arizona, the Galapagos Islands, and Peru. Comparative analyses of cpDNA vs. nuclear data indicate that this history was accompanied by several clear cases of interspecific introgression. Repetitive DNAs contribute roughly half of the total 880 Mb genome, but most transposable element families are relatively old and stable among species. In the genic fraction, pairwise synonymous mutation rates average 1% per my, with non-synonymous changes being about seven times less frequent. Over 1.1 million indels were detected and phylogenetically polarized, revealing a two-fold bias toward deletions over small insertions. We suggest that this genome down-sizing bias counteracts genome size growth by TE amplification and insertions, and helps explain the relatively small genomes that are restricted to this subgenus. Compared to the rate of nucleotide substitution, the rate of indel occurrence is much lower averaging about 17 nucleotide substitutions per indel event.

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

Disciplines
Ecology and Evolutionary Biology | Molecular Genetics | Plant Breeding and Genetics

Comments

Rights
Works produced by employees of the U.S. Government as part of their official duties are not copyrighted within the U.S. The content of this document is not copyrighted.

Authors
Corrinne E. Grover, Mark A. Arick II, Adam Thrash, Justin L. Conover, William S. Sanders, Daniel G. Peterson, James E. Frelighowski, Jodi A. Scheffler, Brian E. Scheffler, and Jonathan F. Wendel

This article is available at Iowa State University Digital Repository: https://lib.dr.iastate.edu/eeob_ag_pubs/321
Insights into the evolution of the New World diploid cottons (*Gossypium*, subgenus *Houzingenia*) based on genome sequencing

Corinne E Grover1*, Mark A Arick II2, Adam Thrash2, Justin L Conover1, William S Sanders2,3,4, Daniel G Peterson2, James E Frelighowski5, Jodi A Scheffler6, Brian E Scheffler7, 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

5USDA, Crop Germplasm Research Unit, College Station, TX, 77845, USA

6USDA, Crop Genetics Research Unit, Stoneville, MS 38776, USA

7USDA, Genomics and Bioinformatics Research Unit, Stoneville, MS 38776, USA

*Author for Correspondence: Corrinne Grover, Ecology Evolution and Organismal Biology, Iowa State University, Ames, IA, USA, 515-294-7098, 515-294-1337, corrinne@iastate.edu

Data deposition: Assembled genomes are available from Genbank under accessions TBD. Raw reads are available through the Short-Read Archive under accessions TBD. Assembled chloroplast genomes are available under Genbank accessions MH477706 through MH477724.
Abstract:

We employed phylogenomic methods to study molecular evolutionary processes and phylogeny in the geographically widely dispersed New World diploid cottons (*Gossypium*, subg. *Houzingenia*). Whole genome resequencing data (average of 33X genomic coverage) were generated to reassess the phylogenetic history of the subgenus and provide a temporal framework for its diversification. Phylogenetic analyses indicate that the subgenus likely originated following trans-oceanic dispersal from Africa about 6.6 mya, but that nearly all of the biodiversity evolved following rapid diversification in the mid-Pleistocene (0.5-2.0 mya), with multiple long-distance dispersals required to account for range expansion to Arizona, the Galapagos Islands, and Peru. Comparative analyses of cpDNA vs. nuclear data indicate that this history was accompanied by several clear cases of interspecific introgression. Repetitive DNAs contribute roughly half of the total 880 Mb genome, but most transposable element families are relatively old and stable among species. In the genic fraction, pairwise synonymous mutation rates average 1% per my, with non-synonymous changes being about seven times less frequent. Over 1.1 million indels were detected and phylogenetically polarized, revealing a two-fold bias toward deletions over small insertions. We suggest that this genome down-sizing bias counteracts genome size growth by TE amplification and insertions, and helps explain the relatively small genomes that are restricted to this subgenus. Compared to the rate of nucleotide substitution, the rate of indel occurrence is much lower averaging about 17 nucleotide substitutions per indel event.

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

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 et al. 1985; Wendel & Grover 2015; Álvarez et al. 2005; Fryxell 1979) (Figure 1). Included in the 13-14 species presently recognized in subgenus *Houzingenia* (Ulloa et al. 2013; Wendel & 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 allopolyplid (AD-genome) cotton, *G. raimondii* Ulbrich (reviewed in (Wendel & 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 their 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 (Cronn et al. 1996; Seelanan et al. 1997; Small & Wendel 2000; Wendel & Albert 1992; Wendel et al. 1995; Álvarez et al. 2005), at least at the sectional 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 & Albert 1992); simple sequence repeat (SSR) and expressed sequence tag (EST)-SSR markers (Zhu et al. 2009; Guo et al. 2007); random amplified polymorphic DNA (RAPD) markers (Khan et al. 2000); internal transcribed sequences (ITS) (Álvarez et al. 2005); and a few single-copy nuclear genes (Álvarez et al. 2005). Relationships among the six subsections remain unclear, with different studies yielding alternative topologies (Cronn et al. 1996; Liu & Wendel 2001; Small & Wendel 2000; Álvarez et al. 2005); however, early morphological and cytogenetic comparisons using intergenomic hybrids have firmly established *G. raimondii* as the closest living relative to the D-genome ancestor of polyploid cotton species (reviewed in (Wendel & Cronn 2003)). Subsequent analyses have supported this observation (Abdalla et al. 2001; Cronn et al. 1999; Liu & Wendel 2001; Cronn et al. 1996; Seelanan et al. 1997; Small et al. 1998; Small & Wendel 2000) with few conflicts (however, see (Wendel et al. 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 (Cronn & 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; Cronn & Wendel 2004; Cronn et al. 2003)). Most remarkably, *G. gossypioides* appears to have been introgressed multiple times, 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 (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 (Álvarez et al. 2005).

Early attempts at understanding the evolution of the repetitive fraction of the genus support the inference of African introgression in *G. gossypioides* (Xin Ping 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 to 934
Mb (Hendrix & James McD Stewart 2005). 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 sequences gained and/or lost 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 (Hawkins et al. 2006; Renny-Byfield et al. 2016; Grover et al. 2017). Furthermore, research comparing the two sister genera to cotton (i.e., *Kokia* and *Gossypioides*; (Grover et al. 2017)) 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; Grover et al. 2017).

Here we reexamine phylogenetic relationships and molecular 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.

**Materials and Methods**

*Sequence generation and initial processing.* DNA was extracted from leaves using either (1) a modified version of the method described by (Dabo et al. 1993), or (2) the Qiagen DNeasy Plant Mini Kit (69104) followed by the DNeasy PowerClean Pro Cleanup kit (12997). For those accessions with sufficient DNA available from USDA-ARS, Stoneville MS (Supplementary Table 3, BGI) samples were submitted to BGI Genomics (Hong Kong) for Illumina library preparation and 2x100bp sequencing. For accessions with limited amounts of available DNA (Supplementary Table 3, NXT), Illumina sequencing libraries were prepared in-house at the USDA-ARS GBRU core facility by the Nextera™ DNA Library Prep Kit (product number FC-121-1030 with adapter set FC-121-1011, Illumina, San Diego, CA, USA) according to the manufacturer’s protocol. Samples obtained from Iowa State University, Ames IA (Supplementary Table 3, USDA) were prepared at the USDA-ARS GBRU core facility using Accel-NGS 2S PCR-Free (Product number 20024 with adapter set 26396, Swift Biosciences, Ann Arbor, MI, USA). Library sizes were validated on the Agilent TapeStation 2200 High Sensitivity D1000 Assay (Part No. 5067-5584, Agilent Technologies, Santa Clara, CA, USA) and assayed for concentration prior to equimolar pooling by a KAPA Library Quantification Kit (Product number KK4854, Kapa Biosystems, Inc, Wilmington, MA, USA) on a qPCR instrument (LightCycler 96, Roche Applied Science, Indianapolis, IN, USA). Each pool was clustered onboard an Illumina HiSeq2500 DNA sequencer with a HiSeq PE (paired-end) Rapid v2 flowcell clustering kit (Product number PE-402-4002, Illumina, San Diego, CA, USA) and sequenced as 2x100 bp with the HiSeq Rapid SBS Kit v2 (Product number FC-402-4021, Illumina, San Diego, CA, USA). The remaining samples (Supplementary Table 3, Novogene) were submitted to Novogene (Beijing) for Illumina library
preparation and 2x150bp sequencing. Reads are available from the Short-Read Archive (SRA) under PRJNA488266. The outgroup, G. longicalyx, was downloaded from SRA (SRX204849) and processed alongside the Houzingenia samples.

Reads were trimmed and filtered with Trimmomatic v0.32 (Anthony M Bolger et al. 2014) 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 (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 & Yandell 2011) using evidence from: (1) the NCBI G. raimondii EST database (Udall et al 2006), (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 (Borodovsky 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), a reference-assisted scaffold, was used to scaffold the selected assemblies against the gold standard G. raimondii genome. MAKER v2.31.6 (Holt & Yandell 2011) 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. Assemblies are also available under PRJNA488266.

**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 & Durbin 2009), post-processed with samtools (Li et al. 2009), and individual genes were independently assembled for each species/accession via BamBam v 1.3 (Page et al. 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 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 (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).

Molecular evolutionary analyses were conducted in R v3.4.4 (Team 2017). Species divergence time estimates were calculated via chronos from {ape} (Paradis et al. 2004), using the divergence estimates previously calculated for the Malvaceae (Grover et al. 2017) and penalized likelihood (Sanderson 2002; Kim & Sanderson 2008) and maximum likelihood. Minimum and maximum node ages were specified for
both the root and the node that separates *Erioxylum* from the rest of the subgenus, using T=dS/r and the minimum/maximum dS for each. Trees were visualized using the (ape) package (Paradis et al. 2004). Ancestral state reconstructions for genome size were completed using fastAnc from (phytools) (Revell 2012). Indels and SNPs were characterized among *Houzingenia* using the Genome Analysis ToolKit (Van der Auwera et al. 2013) and the *G. raimondii* reference sequence (Paterson et al. 2012). SNP introgression was measured both by individual SNP proportions (see https://github.com/IGBB/D_Cottons_USDA/) and ANGSD (Korneliussen et al. 2014). Indel effects on genes were measured by SnpEff (Cingolani, Platts, et al. 2012) and SnpSift (Cingolani, Patel, et al. 2012).

Gene orthology among species was determined via OrthoFinder (Emms & Kelly 2015), and copy numbers per species/gene family was derived from the resulting orthologous clusters. Copy number evolution was modeled using Count (Csurös 2010), which uses a likelihood-based phylogenetic birth-and-death model to estimate gene family sizes along edges and subsequently reconstruct ancestral states. Bootstrap-like replicates were estimated 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 (Bolger et al. 2014) 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 (Hendrix & James Mcd Stewart 2005; Wendel et al. 2002). These 1% GSEs were combined as input into the RepeatExplorer pipeline (Novák et al. 2010, 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 2018 (Grover et al. 2017), which uses the RepeatExplorer implementation of RepeatMasker (Smit et al.) 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 (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* (Grover et al. 2017), 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 (Grover et al. 2017) 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.

Results

Genome assemblies and annotation

Approximately 22-65X raw coverage libraries were sequenced for at least one representative of each D-genome species (Supplementary 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: 67.2 – 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 1). These metrics are comparable to those generated by the subgenus Houzingenia-derived reference 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 1).

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; Cronn et al. 2002). 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 22 *Houzingenia* accessions and for the outgroup, *G. longicalyx* (subgenus *Longiloba*). Maximum likelihood reconstruction of the phylogenetic relationships among species largely recover established section and subsection relationships (Figure 2). 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 2). Species relationships are largely congruent with the most recent phylogenetic inferences for the subgenus using nuclear genes (Álvarez et al. 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 sister to the rest of the arborescent cottons of subsection *Erioxylum* and not sister to the *G. aridum* accession from Jalisco (Álvarez et al. 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 & Wendel 1992). Indeed, phylogenetic analysis of the entire chloroplast for *Houzingenia* species (Figure 3) concurs with previous chloroplast restriction site analysis (Wendel & 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 2014; Ulloa et al. 2006; Feng et al. 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. This was determined using an ABBA-BABA test (Korneliussen et al. 2014; Sousa & Hey 2013) 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 species into *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 2) gives the same results as the ABBA-BABA test (chi-square p-value = 0), 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 2), 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 support the broader conclusion that ancient nuclear introgression occurred 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 supports previous observations of Austroamericana-derived introgression in subsection 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 (Wendel & Albert 1992; Cronn et al. 2003; Cronn & Wendel 2004), and (2) nuclear introgression, as evidenced by the presence of African cotton-like ITS (Wendel et al. 1995) and repetitive DNA (X.-P. 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.

**Recent divergence in subgenus Houzingenia is reflected in the low rate of molecular evolution**

Divergence times were estimated for the thirteen extant Houzingenia species (Figure 2) using the synonymous substitution rate for the Malvaceae, as described in Grover et al (2018). Subgenus Houzingenia diverged an estimated 6.58 mya from the remaining cotton subgenera (represented by Longiloba), a value within prior estimates (Senchina et al. 2003). The lineage leading to G. gossypioides was inferred as the first to diverge from the rest of the subgenus, approximately 2.56 mya (Figure 2), 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 calibrated using the next most basal node, which separates section Erioxylum from the remaining subgenus (see methods), in conjunction with the root. 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 (Wendel & 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 3). While this close relationship between G. davidsonii and G. klotzschianum has been reported previously (Wendel & 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 3). 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 3). Intron 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 earliest-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 ancient divergence of G. gossypioides with the rest of Houzingenia. 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 3).

Transposable elements in *Houzingenia* are older and concordant with small genome sizes

Similar to previous reports (Paterson et al. 2012), 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 4). 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 sampling was insufficient to generate a confidence ellipse (i.e., *Selera* and *Integrifolia*), the plotted data points are contained within the occupied space of another subsection (Figure 4, inset). *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 4).

To compare the overlap among subsections, we performed a Procrustes ANOVA, as implemented in the R package (geomorph) (Adams & Otárola-Castillo 2013). 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 4). 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 (Hawkins et al. 2006, 2009; Grover et al. 2007). 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 (Supplementary Figure 1), 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) 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* (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, demonstrates both amplification and removal concomitant with the inferred changes in overall genome size (Figure 2; 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 (Britten et al. 2003; Tuğrul et al. 2015; Lin et al. 2017; Halligan et al. 2013). 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 (Kapusta et al. 2017; Chintalapati et al. 2017; Sato et al. 2012). 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) reproduces 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 smallest and longest 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 5).

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 2), 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; Supplementary Table 2). 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 2). 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 earliest-diverging lineage, *G. gossypioides*, consistently has more indels relative to SNPs, possibly as a consequence of its introgressed history (Wendel et al. 1995; Cronn et al. 2003). 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 at this scale; however,
these results do support the idea that small deletions may be able to partially counteract genome size growth by TE amplification and small insertions.

The genic consequences for these indels were evaluated for the 37,223 gene models in the *G. raimondii* reference (Paterson et al. 2012). Less than 1.5% of indels (15,786) had any in-gene effects in any species, of which 12,679 (19%) only result in a single amino acid gain or loss (1,333 and 1,663 indels, respectively). Nearly 50% of exonic indels resulted in a frameshift mutation, 8% of which had additional consequences (e.g., gain or loss of start, stop, or splice signal). Over 27% affected the protein length only, with a slight bias (2:1) toward inframe deletions and only 1.4% of these affecting the start or stop codons.

In total, 9342 genes were affected by indels in at least one species; however, most species exhibited indel-induced genic changes in an average of 2700 genes, of which approximately 600 induce length changes only. Notably, while the *G. raimondii* accession sequenced had the fewest indels detected in genes, 1.8% of the gene models were nevertheless affected in this accession. Given the relative uniformity of *G. raimondii* in protein-coding sequences (Wendel, unpubl.), this may represent the amount of error inherent in the indel analysis due to the bioinformatic identification of indels or to the gene models represented in the published genome.

**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 (Golicz, Batley, et al. 2016; Golicz, Bayer, et al. 2016; Montenegro et al. 2017; Hirsch et al. 2014; Lai et al. 2010; Ying-hui Li et al. 2014; Lin et al. 2014; Pinosio et al. 2016; Schatz et al. 2014). 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 (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 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 5). 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 representing all species in the subgenus to characterize the evolutionary history of the subgenus and provide insight into the molecular evolution among closely related species.

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 previously established sectional and subsection relationships (Figure 2), and 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 arose from the earliest diverging lineage within the clade, as suggested by earlier work using rather limited genomic sampling. This is a rare, highly localized species from Oaxaca, Mexico, 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) remains 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 & 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.

Our estimates of initial divergence for the subgenus are consistent with previous estimates from the chloroplast genome (Senchina et al. 2003), and we also find that while the subgenus appears to have originated about 6.6 mya, all surviving species trace to a much more recent origin in the Pleistocene.
(about 2.5 mya). 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 (Wendel & Grover 2015).

**Phylogenetic incongruence and ancient hybridization.**

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 3). Part of the reason for this may be a history of documented (e.g., G. aridum, G. gossypioides) as well as unobserved interspecific introgression and rapid radiation at the base of the clade, which generates short (i.e., difficult to resolve) internodes. These results recapitulate some of our earlier work (Alvarez & Wendel 2006; Álvarez et al. 2005; Wendel & Albert 1992; Cronn & Wendel 2004; Wendel et al. 1995; Cronn et al. 2003) in which we highlight how comparison between nuclear vs. chloroplast phylogenies may inform ancient hybridization events, e.g., the evolutionary histories of G. aridum and G. gossypioides. Populations of the wide-ranging G. aridum from the single Mexican state of Colima, for example, 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) (Alvarez & Wendel 2006; Wendel & Albert 1992). We obtained this same incongruence in our analysis, with the added twist that in the reconstruction based on the nuclear genome, G. aridum from Colima appears as the sister to the rest of 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 (Wendel & Grover 2015)), or if instead G. aridum from Colima was “dragged” to its early-diverging 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 (Alvarez & Wendel 2006) 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 2), 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.

Likewise, comparison between the nuclear and chloroplast phylogenies (Figure 3) reveals the previously observed striking incongruence between the nuclear and cpDNA placement of G. gossypioides. As described before (X.-P. Zhao et al. 1998)(Wendel & Albert 1992; Cronn et al. 2003; Cronn & Wendel 2004), G. gossypioides is recovered as sister to 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 3) 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 (Brown & Menzel 1952; Menzel & Brown 1955)). 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 (Cronn et al. 2003; Wendel et al. 1995; Cronn et al. 1996; Zhao et al. 1995). 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 (Xin Ping Zhao et al. 1998). Although we did not observe introgression 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.

**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.

**Protein evolution.** With respect to genic evolution, we report a relatively narrow range of interspecific non-synonymous substitution rate (dN), averaging 0.0014 non-synonymous substitutions per site per million years, with a synonymous substitution rate about an order of magnitude higher (Table 3). Thus, evolution at the amino acid level is inferred to be quite slow, averaging only about 1% per codon every 7 million years. 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 (Gaut et al. 2011; Smith & Donoghue 2008). Interestingly, however, indels were estimated to affect as many as one quarter of the gene models in at least one species, with an average of 7% per non-reference species (compared to 1.8% for G. raimondii acc. 8 compared to the G. raimondii-derived reference genome). Together with the estimates of copy number variability (see Results), these results warrant a closer inspection on the evolution of genes and gene content in these species.

**Transposable elements and the repetitive fraction.** Similar to previous reports for Gossypium (Paterson et al. 2012; Yuan et al. 2015; Fuguang Li et al. 2014; Yu et al. 2012; Wang et al. 2012; Zhang et al. 2015; Li et al. 2015), 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 4). Relatively few repetitive elements differ significantly in
copy number among the species (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 4) and elsewhere in the genus (Hawkins et al. 2006) 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 (Estep et al. 2013; Stein et al. 2018; Daron et al. 2014; Luo et al. 2017; Wang & Dooner 2006). 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 (Supplementary Figure 1). 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.

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*). While *copia* elements comprise a higher proportion of the genome for *Houzingenia* species than for other cottons surveyed (Supplementary Figure 5), these elements generally seem to be in decline (Table 4), 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 (Kapusta et al. 2017; Chintalapati et al. 2017; Sato et al. 2012; Stein et al. 2018). Despite relatively recent divergence times, we found over one million positions associated with an indel in at least one of the 13 *Houzingenia* species (relative to *Longiloba*), a third of which distinguish one or more *Houzingenia* species. Although indels were found genome-wide, there was variation among chromosomes, which ranged 1.5-fold per Mb. Most indels were small, averaging 6.2 nucleotides, with a range in size of 1 – 270 nucleotides (Table 5). 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 (Supplementary Table 2). 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 (Michael 2014; Simonin & Roddy 2018; Hu et al. 2010; Grover & Wendel 2010). 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 the finest scale of aligned nucleotides, a dynamic process of genomic downsizing that was inferred from computational modeling a decade ago (Hawkins et al. 2009).

Acknowledgements
We thank Li Wang and Matthew Hufford for discussion relating to introgression; Tracy Heath and Walker C Pett for insight into phylogenetic modeling of gene copy number evolution; Mary Duke for assistance in data generation; and the Iowa State University ResearchIT Unit (http://researchit.las.iastate.edu/) for computational support. Funding for this project came from the NSF Plant Genome Program (141589) to JW, Cotton Incorporated (13-600) to BES, and the USDA ARS (6066-21310-004-00-D, BES; 6066-21000-052-00D, JAS; 3091-21000-041-00-D, JEF; 6066-21310-004-23-S, DP).

References


Chen Z et al. 2016. Chloroplast DNA structural variation, phylogeny, and age of divergence among


Estep MC, Debarry JD, Bennetzen JL. 2013. The dynamics of LTR retrotransposon accumulation across 25


Li F et al. 2014. Genome sequence of the cultivated cotton gossypium arboreum. Nat. Genet. 46. doi: 10.1038/ng.2987.


Yuan D et al. 2015. The genome sequence of Sea-Island cotton (Gossypium barbadense) provides insights into the allopolyploidization and development of superior spinnable fibres. Sci. Rep. 5. doi: 10.1038/srep17662.


Zhao XP et al. 1998. Dispersed repetitive DNA has spread to new genomes since polyploid formation in cotton. Genome Res. 8:479–492. doi: 10.1101/gr.8.5.479.


**Figure legends**

Figure 1: Approximate geographic ranges of Houzingenia species. 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 2: 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 in the methods.

Figure 3: 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 position of G. aridum Colima on the nuclear phylogeny (right) has been added to the figure with a dotted line, as presence of this accession “attracts” G. schwendimanii to its position thereby distorting the topology; the non-introgressed topology is pictured here. The chloroplast phylogeny shown here was derived from WGS-derived whole chloroplast sequences (see methods); this tree topology was also recovered from a concatenated chloroplast gene-only phylogenetic analysis 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 supported between the trees (in colors), the relationship among subsections varies between the two molecule types.

Figure 4: 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. Repetitive clusters that did not match any sequence in the database are denoted by the asterisk, whereas repetitive clusters that matched conflicting categories were classified as “Unspecified”. Inset: 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). Confidence intervals (95%) are shown for subgenera with a minimum of four representatives.

Figure 5: 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: 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.

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).

Supplementary Figure 4: 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.
A. Rate of gain, per million years

B. Rate of loss, per million years
## Scaffold Statistics

<table>
<thead>
<tr>
<th>Subsection</th>
<th>Species</th>
<th>Accession</th>
<th># contigs (&gt;= 1kb)</th>
<th>Largest contig (Mb)</th>
<th>Contigs (&gt;= 1 kb)</th>
<th>Contigs (&gt;= 25 kb)</th>
<th>Contigs (&gt;= 50kb)</th>
<th>Total length</th>
<th>Genome Size (GS)</th>
<th>% Genome Covered</th>
<th>N50</th>
<th>% N</th>
<th># Gene Models</th>
<th># BUSCOs</th>
<th>% recovered</th>
<th>% partial BUSCO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austroamericana</td>
<td>G. raimondii</td>
<td>Paterson 2012</td>
<td>1,033</td>
<td>70.71</td>
<td>761.41</td>
<td>754.80</td>
<td>753.82</td>
<td>761.41</td>
<td>880</td>
<td>86.5%</td>
<td>62.18</td>
<td>2%</td>
<td>880</td>
<td>30,475</td>
<td>93%</td>
<td>27</td>
</tr>
<tr>
<td>Caducibracteata</td>
<td>G. armourianum</td>
<td>D2-1-6</td>
<td>13,359</td>
<td>55.08</td>
<td>645.03</td>
<td>600.41</td>
<td>599.97</td>
<td>671.70</td>
<td>856</td>
<td>78.5%</td>
<td>47.59</td>
<td>15%</td>
<td>35,142</td>
<td>1,285</td>
<td>89%</td>
<td>47</td>
</tr>
<tr>
<td>Caducibracteata</td>
<td>G. harknessii</td>
<td>JFW</td>
<td>20,602</td>
<td>52.42</td>
<td>615.99</td>
<td>541.69</td>
<td>540.67</td>
<td>643.05</td>
<td>910</td>
<td>70.7%</td>
<td>43.64</td>
<td>7%</td>
<td>36,068</td>
<td>1,294</td>
<td>90%</td>
<td>51</td>
</tr>
<tr>
<td>Caducibracteata</td>
<td>G. turneri</td>
<td>D10-7</td>
<td>18,841</td>
<td>49.15</td>
<td>742.21</td>
<td>654.97</td>
<td>601.54</td>
<td>774.62</td>
<td>910</td>
<td>85.1%</td>
<td>33.33</td>
<td>2%</td>
<td>45,244</td>
<td>1,366</td>
<td>95%</td>
<td>17</td>
</tr>
<tr>
<td>Erioxylum</td>
<td>G. aridum</td>
<td>DRD-185</td>
<td>21,813</td>
<td>52.48</td>
<td>619.71</td>
<td>552.02</td>
<td>551.45</td>
<td>648.51</td>
<td>919</td>
<td>70.6%</td>
<td>42.90</td>
<td>8%</td>
<td>35,142</td>
<td>1,285</td>
<td>89%</td>
<td>47</td>
</tr>
<tr>
<td>Erioxylum</td>
<td>G. lobatum</td>
<td>D7-157</td>
<td>22,383</td>
<td>53.83</td>
<td>625.55</td>
<td>555.17</td>
<td>554.44</td>
<td>654.85</td>
<td>934</td>
<td>70.1%</td>
<td>43.64</td>
<td>8%</td>
<td>35,572</td>
<td>1,310</td>
<td>91%</td>
<td>41</td>
</tr>
<tr>
<td>Erioxylum</td>
<td>G. laxum</td>
<td>D9-4</td>
<td>16,668</td>
<td>60.54</td>
<td>689.25</td>
<td>623.49</td>
<td>621.55</td>
<td>720.31</td>
<td>934</td>
<td>77.1%</td>
<td>48.55</td>
<td>13%</td>
<td>32,375</td>
<td>1,321</td>
<td>92%</td>
<td>37</td>
</tr>
<tr>
<td>Erioxylum</td>
<td>G. schwendimanii</td>
<td>D11-1</td>
<td>18,906</td>
<td>52.32</td>
<td>623.31</td>
<td>526.16</td>
<td>513.54</td>
<td>651.41</td>
<td>929</td>
<td>70.1%</td>
<td>40.29</td>
<td>5%</td>
<td>38,314</td>
<td>1,348</td>
<td>94%</td>
<td>25</td>
</tr>
<tr>
<td>Houzingenia</td>
<td>G. thurberi</td>
<td>D1-35</td>
<td>15,309</td>
<td>47.60</td>
<td>582.19</td>
<td>505.27</td>
<td>498.80</td>
<td>605.21</td>
<td>841</td>
<td>72.0%</td>
<td>37.86</td>
<td>4%</td>
<td>37,553</td>
<td>1,342</td>
<td>93%</td>
<td>26</td>
</tr>
<tr>
<td>Houzingenia</td>
<td>G. trilobum</td>
<td>D8-8</td>
<td>14,099</td>
<td>44.53</td>
<td>562.98</td>
<td>483.41</td>
<td>474.98</td>
<td>586.05</td>
<td>851</td>
<td>68.9%</td>
<td>36.15</td>
<td>5%</td>
<td>36,663</td>
<td>1,321</td>
<td>92%</td>
<td>43</td>
</tr>
<tr>
<td>Integrifolia</td>
<td>G. davidsonii</td>
<td>D3D-27</td>
<td>16,779</td>
<td>48.06</td>
<td>603.64</td>
<td>517.19</td>
<td>506.66</td>
<td>629.89</td>
<td>910</td>
<td>69.2%</td>
<td>38.65</td>
<td>3%</td>
<td>38,755</td>
<td>1,237</td>
<td>86%</td>
<td>41</td>
</tr>
<tr>
<td>Integrifolia</td>
<td>G. hortischianum</td>
<td>D3K-57</td>
<td>16,881</td>
<td>46.26</td>
<td>569.35</td>
<td>495.76</td>
<td>492.77</td>
<td>596.12</td>
<td>880</td>
<td>67.7%</td>
<td>37.84</td>
<td>5%</td>
<td>37,444</td>
<td>1,332</td>
<td>93%</td>
<td>37</td>
</tr>
<tr>
<td>Selero</td>
<td>G. gossypioides</td>
<td>D6-5</td>
<td>23,734</td>
<td>42.78</td>
<td>554.55</td>
<td>446.85</td>
<td>441.62</td>
<td>585.41</td>
<td>841</td>
<td>69.6%</td>
<td>33.53</td>
<td>4%</td>
<td>26,492</td>
<td>1,133</td>
<td>79%</td>
<td>41</td>
</tr>
</tbody>
</table>
Table 2: Number of shared, derived nuclear SNPs between *G. davidsonii* (Integrifolia) and *G. aridum* accessions from Colima and Jalisco. Previous research indicates that Colima *G. aridum* has Integrifolia-derived cytoplasm and nuclear sequences. *G. gossypioides* was used for ancestral states.

<table>
<thead>
<tr>
<th></th>
<th><em>G. aridum</em> (Colima)</th>
<th><em>G. aridum</em> (Jalisco)</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>overall SNPs</td>
<td>188,472</td>
<td>182,563</td>
<td>0.0005</td>
</tr>
<tr>
<td>Chr01 SNPs</td>
<td>12,808</td>
<td>12,808</td>
<td>1.0000</td>
</tr>
<tr>
<td>Chr02 SNPs</td>
<td>17,118</td>
<td>17,094</td>
<td>0.8941</td>
</tr>
<tr>
<td>Chr03 SNPs</td>
<td>11,956</td>
<td>11,353</td>
<td>0.0005</td>
</tr>
<tr>
<td>Chr04 SNPs</td>
<td>17,292</td>
<td>16,643</td>
<td>0.0005</td>
</tr>
<tr>
<td>Chr05 SNPs</td>
<td>18,950</td>
<td>18,065</td>
<td>0.0005</td>
</tr>
<tr>
<td>Chr06 SNPs</td>
<td>11,013</td>
<td>10,732</td>
<td>0.0600</td>
</tr>
<tr>
<td>Chr07 SNPs</td>
<td>15,822</td>
<td>14,649</td>
<td>0.0005</td>
</tr>
<tr>
<td>Chr08 SNPs</td>
<td>12,904</td>
<td>12,911</td>
<td>0.9795</td>
</tr>
<tr>
<td>Chr09 SNPs</td>
<td>15,131</td>
<td>14,922</td>
<td>0.2399</td>
</tr>
<tr>
<td>Chr10 SNPs</td>
<td>17,585</td>
<td>16,895</td>
<td>0.0005</td>
</tr>
<tr>
<td>Chr11 SNPs</td>
<td>15,741</td>
<td>14,941</td>
<td>0.0005</td>
</tr>
<tr>
<td>Chr12 SNPs</td>
<td>8,600</td>
<td>8,636</td>
<td>0.8081</td>
</tr>
<tr>
<td>Chr13 SNPs</td>
<td>13,552</td>
<td>12,914</td>
<td>0.0005</td>
</tr>
<tr>
<td>genic SNPs</td>
<td>7,843</td>
<td>7,419</td>
<td>0.0015</td>
</tr>
<tr>
<td>number of genes</td>
<td>4,808</td>
<td>4,721</td>
<td>0.3733</td>
</tr>
</tbody>
</table>
Table 3: Median synonymous (bottom) and non-synonymous (top) mutation rates between species. Values in parentheses represent upper- and lower quartile, respectively.

<table>
<thead>
<tr>
<th>Species</th>
<th>Synonymous Rate</th>
<th>Non-synonymous Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Houzingenia</td>
<td>0.0013-0.0049</td>
<td>0.0011-0.0045</td>
</tr>
<tr>
<td>G. gossypioides</td>
<td>0.0020</td>
<td>0.0012-0.0048</td>
</tr>
<tr>
<td>G. klotzschianum</td>
<td>0.0021</td>
<td>0.0012-0.0047</td>
</tr>
<tr>
<td>G. integrifolia</td>
<td>0.0020</td>
<td>0.0012-0.0049</td>
</tr>
<tr>
<td>G. trilobum</td>
<td>0.0023</td>
<td>0.0013-0.0051</td>
</tr>
<tr>
<td>G. schwendemanii</td>
<td>0.0025</td>
<td>0.0016-0.0055</td>
</tr>
<tr>
<td>Erioxylum</td>
<td>0.0021</td>
<td>0.0013-0.0048</td>
</tr>
<tr>
<td>G. lobatum</td>
<td>0.0022</td>
<td>0.0012-0.0049</td>
</tr>
<tr>
<td>G. turneri</td>
<td>0.0024</td>
<td>0.0013-0.0050</td>
</tr>
<tr>
<td>Caducibracteata</td>
<td>0.0025</td>
<td>0.0016-0.0055</td>
</tr>
<tr>
<td>Houzingenia</td>
<td>0.0021</td>
<td>0.0013-0.0048</td>
</tr>
<tr>
<td>G. gossypioides</td>
<td>0.0025</td>
<td>0.0013-0.0048</td>
</tr>
<tr>
<td>G. klotzschianum</td>
<td>0.0030</td>
<td>0.0015-0.0052</td>
</tr>
<tr>
<td>G. integrifolia</td>
<td>0.0032</td>
<td>0.0015-0.0052</td>
</tr>
<tr>
<td>G. trilobum</td>
<td>0.0032</td>
<td>0.0015-0.0052</td>
</tr>
<tr>
<td>G. schwendemanii</td>
<td>0.0030</td>
<td>0.0015-0.0052</td>
</tr>
<tr>
<td>Erioxylum</td>
<td>0.0030</td>
<td>0.0015-0.0052</td>
</tr>
<tr>
<td>G. lobatum</td>
<td>0.0030</td>
<td>0.0015-0.0052</td>
</tr>
<tr>
<td>G. turneri</td>
<td>0.0030</td>
<td>0.0015-0.0052</td>
</tr>
<tr>
<td>Caducibracteata</td>
<td>0.0030</td>
<td>0.0015-0.0052</td>
</tr>
</tbody>
</table>
Table 4: Gain and loss in *copia* elements for each of the accessions clusters

<table>
<thead>
<tr>
<th>Species</th>
<th>Accession</th>
<th>Numbers of clusters with gain or loss, per accession</th>
<th>Sequence loss in <em>copia</em> elements (Mb)</th>
<th>Sequence gain in <em>copia</em> elements (Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>G. raimondii</em></td>
<td>Paterson 2012</td>
<td>28 ↓ 15 ↑</td>
<td>-7.1</td>
<td>2.4</td>
</tr>
<tr>
<td><em>G. raimondii</em></td>
<td>acc 2</td>
<td>22 ↓ 21 ↑</td>
<td>-7.6</td>
<td>8.6</td>
</tr>
<tr>
<td><em>G. raimondii</em></td>
<td>acc 31</td>
<td>19 ↓ 24 ↑</td>
<td>-6.8</td>
<td>2.4</td>
</tr>
<tr>
<td><em>G. raimondii</em></td>
<td>acc 4</td>
<td>27 ↓ 16 ↑</td>
<td>-10.3</td>
<td>5.6</td>
</tr>
<tr>
<td><em>G. raimondii</em></td>
<td>acc 53</td>
<td>29 ↓ 14 ↑</td>
<td>-11.9</td>
<td>3.4</td>
</tr>
<tr>
<td><em>G. raimondii</em></td>
<td>acc 6</td>
<td>13 ↓ 30 ↑</td>
<td>-1.4</td>
<td>20.2</td>
</tr>
<tr>
<td><em>G. raimondii</em></td>
<td>acc 8</td>
<td>28 ↓ 15 ↑</td>
<td>-7.0</td>
<td>4.9</td>
</tr>
<tr>
<td><em>G. armourianum</em></td>
<td>acc 6</td>
<td>23 ↓ 20 ↑</td>
<td>-4.3</td>
<td>6.1</td>
</tr>
<tr>
<td><em>G. harknessii</em></td>
<td>acc 2</td>
<td>30 ↓ 13 ↑</td>
<td>-9.0</td>
<td>3.9</td>
</tr>
<tr>
<td><em>G. turneri</em></td>
<td>acc 3</td>
<td>20 ↓ 23 ↑</td>
<td>-6.8</td>
<td>6.6</td>
</tr>
<tr>
<td><em>G. turneri</em></td>
<td>acc 7</td>
<td>30 ↓ 13 ↑</td>
<td>-9.2</td>
<td>4.7</td>
</tr>
<tr>
<td><em>G. turneri</em></td>
<td>acc 8</td>
<td>27 ↓ 16 ↑</td>
<td>-7.5</td>
<td>4.2</td>
</tr>
<tr>
<td><em>G. aridum</em></td>
<td>acc 185</td>
<td>23 ↓ 20 ↑</td>
<td>-7.8</td>
<td>5.4</td>
</tr>
<tr>
<td><em>G. lobatum</em></td>
<td>acc 157</td>
<td>26 ↓ 17 ↑</td>
<td>-9.2</td>
<td>5.2</td>
</tr>
<tr>
<td><em>G. lobatum</em></td>
<td>acc 4</td>
<td>20 ↓ 23 ↑</td>
<td>-3.8</td>
<td>6.1</td>
</tr>
<tr>
<td><em>G. laxum</em></td>
<td>acc 4</td>
<td>20 ↓ 23 ↑</td>
<td>-3.1</td>
<td>4.8</td>
</tr>
<tr>
<td><em>G. schwendimanii</em></td>
<td>acc 1</td>
<td>26 ↓ 17 ↑</td>
<td>-5.6</td>
<td>3.4</td>
</tr>
<tr>
<td><em>G. thurberi</em></td>
<td>acc 2</td>
<td>27 ↓ 16 ↑</td>
<td>-5.0</td>
<td>7.0</td>
</tr>
<tr>
<td><em>G. thurberi</em></td>
<td>acc 35</td>
<td>25 ↓ 18 ↑</td>
<td>-5.1</td>
<td>1.2</td>
</tr>
<tr>
<td><em>G. trilobum</em></td>
<td>acc 8</td>
<td>21 ↓ 22 ↑</td>
<td>-2.3</td>
<td>6.5</td>
</tr>
<tr>
<td><em>G. trilobum</em></td>
<td>acc 9</td>
<td>21 ↓ 22 ↑</td>
<td>-5.6</td>
<td>5.8</td>
</tr>
<tr>
<td><em>G. davidsonii</em></td>
<td>acc 27</td>
<td>22 ↓ 21 ↑</td>
<td>-4.8</td>
<td>3.2</td>
</tr>
<tr>
<td><em>G. klotzschianum</em></td>
<td>acc 56</td>
<td>25 ↓ 18 ↑</td>
<td>-7.2</td>
<td>2.5</td>
</tr>
<tr>
<td><em>G. klotzschianum</em></td>
<td>acc 57</td>
<td>24 ↓ 19 ↑</td>
<td>-3.3</td>
<td>5.0</td>
</tr>
<tr>
<td><em>G. gossypioides</em></td>
<td>acc 5</td>
<td>24 ↓ 19 ↑</td>
<td>-5.4</td>
<td>2.0</td>
</tr>
<tr>
<td><em>G. gossypioides</em></td>
<td>acc 7</td>
<td>26 ↓ 17 ↑</td>
<td>-7.9</td>
<td>5.8</td>
</tr>
<tr>
<td>Chromosome</td>
<td># Indels</td>
<td>Chromosome Length (Mb)</td>
<td># indels/Mb</td>
<td>Average Indel Size (nt)</td>
</tr>
<tr>
<td>------------</td>
<td>---------</td>
<td>-----------------------</td>
<td>-------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>1</td>
<td>63,848</td>
<td>55.9</td>
<td>1,143</td>
<td>6.3</td>
</tr>
<tr>
<td>2</td>
<td>60,823</td>
<td>62.8</td>
<td>969</td>
<td>5.9</td>
</tr>
<tr>
<td>3</td>
<td>48,607</td>
<td>45.8</td>
<td>1,062</td>
<td>6.1</td>
</tr>
<tr>
<td>4</td>
<td>58,550</td>
<td>62.2</td>
<td>942</td>
<td>6.0</td>
</tr>
<tr>
<td>5</td>
<td>49,943</td>
<td>64.1</td>
<td>779</td>
<td>5.7</td>
</tr>
<tr>
<td>6</td>
<td>58,156</td>
<td>51.1</td>
<td>1,139</td>
<td>6.5</td>
</tr>
<tr>
<td>7</td>
<td>67,740</td>
<td>61.0</td>
<td>1,111</td>
<td>6.4</td>
</tr>
<tr>
<td>8</td>
<td>67,069</td>
<td>57.1</td>
<td>1,174</td>
<td>6.5</td>
</tr>
<tr>
<td>9</td>
<td>72,303</td>
<td>70.7</td>
<td>1,022</td>
<td>6.7</td>
</tr>
<tr>
<td>10</td>
<td>59,521</td>
<td>62.2</td>
<td>957</td>
<td>6.0</td>
</tr>
<tr>
<td>11</td>
<td>62,707</td>
<td>62.7</td>
<td>1,000</td>
<td>6.1</td>
</tr>
<tr>
<td>12</td>
<td>40,747</td>
<td>35.4</td>
<td>1,150</td>
<td>6.0</td>
</tr>
<tr>
<td>13</td>
<td>51,732</td>
<td>58.3</td>
<td>887</td>
<td>6.0</td>
</tr>
</tbody>
</table>

Table 5: Indels in *Houzingenia* relative to the outgroup *G. longicalyx* (*Longiloba*), partitioned by chromosome and by species.

<table>
<thead>
<tr>
<th>Species</th>
<th>Chromosome</th>
<th># SNPs</th>
<th># Indels</th>
<th>SNPs:Indels</th>
</tr>
</thead>
<tbody>
<tr>
<td>G. raimondii</td>
<td>D5-8</td>
<td>7,909,366</td>
<td>451,713</td>
<td>18</td>
</tr>
<tr>
<td>G. armourianum</td>
<td>D2-1-6</td>
<td>7,525,371</td>
<td>442,985</td>
<td>17</td>
</tr>
<tr>
<td>G. harknessii</td>
<td>D2-2</td>
<td>8,140,633</td>
<td>474,421</td>
<td>17</td>
</tr>
<tr>
<td>G. turneri</td>
<td>D10-7</td>
<td>8,155,064</td>
<td>475,161</td>
<td>17</td>
</tr>
<tr>
<td>G. aridum</td>
<td>D4-185</td>
<td>8,555,662</td>
<td>487,561</td>
<td>18</td>
</tr>
<tr>
<td>G. lobatum</td>
<td>D7-157</td>
<td>8,651,866</td>
<td>490,322</td>
<td>18</td>
</tr>
<tr>
<td>G. laxum</td>
<td>D9-4</td>
<td>8,015,127</td>
<td>462,728</td>
<td>17</td>
</tr>
<tr>
<td>G. schwendimanii</td>
<td>D11-1</td>
<td>8,606,096</td>
<td>491,961</td>
<td>17</td>
</tr>
<tr>
<td>G. thurberi</td>
<td>D1-35</td>
<td>8,139,420</td>
<td>478,238</td>
<td>17</td>
</tr>
<tr>
<td>G. trilobum</td>
<td>D8-8</td>
<td>8,232,774</td>
<td>482,728</td>
<td>17</td>
</tr>
<tr>
<td>G. davidsonii</td>
<td>D3D-27</td>
<td>8,539,202</td>
<td>493,939</td>
<td>17</td>
</tr>
<tr>
<td>G. klotzschianum</td>
<td>D3K-57</td>
<td>8,545,127</td>
<td>494,072</td>
<td>17</td>
</tr>
<tr>
<td>G. gossypioides</td>
<td>D6-5</td>
<td>8,359,287</td>
<td>513,538</td>
<td>16</td>
</tr>
</tbody>
</table>