RESEARCH

# Thirteen genome sequences representing the entire subgenus Houzingenia (Gossypium): insights into evolution of the New World diploid cottons

Corrinne E Grover[1*†], Mark A Arick II[2], Justin C Conover[1], Jodi A Scheffler[3], William S Sanders[2], Daniel G Peterson[2], Brian E Scheffler[3] and Jonathan F Wendel[1]

[*]Correspondence: corrinne@iastate.edu
[1]Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA 50011, USA
Full list of author information is available at the end of the article
[†]Equal contributor

**Abstract**

**Background:** Text for this section.

**Results:** Text for this section.

**Conclusions:** Text for this section.

**Keywords:** genome sequence; cotton; *Gossypium*; molecular evolution

## 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 Corrinne. Among the 13 species currently included in the D-genome Corrinne areG. *harknessii* (D2-2), an important species for cytoplasmic male sterility in cotton, and *G. raimondii* (D5), the model diploid progenitor to wild and domesticated allopolyploid cotton Corrinne. The close relationship of *Houzingenia* species to the agronomically important polyploids, combined with the relative ease of sampling this subgenus for early cotton taxonomists, facilitated much of the current understanding of the relationships among D-genome species.

These early taxonomists divided subgenus *Houzingenia* into two sections and six subsections, whose species alliances have largely been retained by subsequent phylogenetic studies Corrinne. Several molecular datasets have been used to evaluate these relationships, including chloroplast restriction sites Corrinne; simple sequence repeat (SSR) and expressed sequence tag (EST)-SSR markers Corrinne; random amplified polymorphic DNA (RAPD) markers Corrinne; internal transcribed sequences (ITS) Corrinne; and few single-copy nuclear genes Corrinne. Relationships among the six subsections, however, remain unclear despite numerous, and often conflicting, studies Corrinne. Determining the closest living relative of the D-genome ancestor to the polyploid, however, has been met with greater success. Early morphological and cytogenetic comparisons using intergenomic hybrids quickly identified *G. raimondii* as the closest living relative to the D-genome ancestor of polyploid cotton species Corrinne. Subsequent analyses have largely supported this observation (Abdalla et al., 2001; cronn 1999, liu 2001, Cronn et al., 1996 Seelanan et al., 1997 Small et al., 1998; Small and Wendel, 2000a,b), with few conflicts Corrinne.

Corrinne: which c
tions here, Endriz
et al 1985

Corrinne: althoug
see Ulloa 2013

Corrinne: Endrizz
1985, others?

Corrinne: Cronn
al., 1996; Seelana
et al., 1997; Smal
and Wendel, 200
Wendel and Albe
1992; Wendel et a
1995b; alvarez 20

Corrinne: citation

Corrinne: citation

Corrinne: citation

Corrinne: citation

Corrinne: citation

Corrinne: Cronn
al., 1996; Liu et a
2001b; Small and
Wendel, 2000b m
citations

Corrinne: Stephe
1944b Hutchinson
et al., 1945, 1947
Gerstel and Phill

A secondary outcome of this research has been the elucidation of multiple instances of hybridization among D-genome (i.e, *Houzingenia*) species Corrinne, and, in one remarkable case (i.e., *G. gossypioides*), between a *Houzingenia* species and another, geographically isolated subgenus from Africa (either A-, B-, or, F-genome Corrinne). Notably, *G. gossypioides* is multiply introgressant, with subsequent hybridization to a member of the *G. raimondii* lineage resulting in chloroplast, if not further (and cryptic), nuclear introgression (Cronn 2003, cryptic trysts). Cytoplasmic introgression, and possibly cryptic nuclear, is also present in some populations of *G. aridum*, i.e., the Mexican Colima populations; *G. aridum* accessions derived from this location possess a *G. davidsonii*- or *G. klotzschianum*-like cytoplasm.

Corrinne: cite Cryptic Trysts

Corrinne: cite cryptic Trysts again

Modest attempts at understanding the evolution of the repetitive fraction of this genus support the inference of African introgression in *G. gossypioides* Corrinne; however, little else is understood with respect to the evolution of the non-genic fraction of *Houzingenia*. The D-genome cottons possess the smallest genome sizes in the genus, ranging only  1.11 fold, from 841 Mb – 934 Mb. Notably, the distribution of genome sizes among the subsections suggests that this subgenus has experienced differential growth and/or reduction in genome size among species Corrinne; however, the patterns of sequence gain and loss have not been characterized for the subgenus. While the differences in genome size are not dramatic, there is evidence that the transposable element types which have accumulated in *G. raimondii* are different than those that have achieved higher copy numbers than the remainder of the genus Corrinne. Furthermore, research comparing the two sister genera to cotton (i.e., *Kokia* and *Gossypioides*; Corrinne) reveals that their apparently static genome sizes belies both gain and loss of repetitive sequence, a result similar to that of the extant members of the A-genome (subgenus *Gossypium*), whose small change in genome size ( 1.05X) masks differences in element accumulation Corrinne.

Corrinne: Zhao 1998

Corrinne: need a figure

Corrinne: Jennifer, A domestication, kokia

Corrinne: cite Kokia paper

Corrinne: cite Agenome domestication

Modern sequencing techniques make it easy to produce a substantial amount of genomic sequencing suitable for addressing these basic questions in a more genomically comprehensive manner. Here we use modest coverage Illumina sequencing to present an in-depth view of the evolution subgenus Houzingenia, the cotton D-genome clade. We leverage newly generated genome and plastome sequences, representing the first for many species, to address questions surrounding genome evolution in a monophyletic group of closely related species. We characterize the patterns of molecular evolution of both genes and repetitive sequences to provide insight into the pace and pattern of evolution in this subgenus. For the first time, intergenic regions are evaluated to characterize the amount of divergence outside of genes, and due to indels or single-nucleotide polymorphisms (SNPs). Finally, we revisit the phylogeny of the D-genome, both adding additional insight into the relationships among species using hundreds of nuclear genes, as well as addressing questions regarding sequence gain and loss among closely related species. The genome characterized here not only provides insight into molecular evolution on a relatively recent timeframe, but it also provides resources for comparative research and the cotton community at large.

## Results

### Genome assemblies and annotation

Approximately 20X raw coverage libraries Corrinne were sequenced for at least one

Corrinne: update numbers

representative of each D-genome species (Table WHATEVER), resulting in an average of 54 M Corrinne reads per accession. Quality filters further reduced the number of reads per sample to an average of 38 M (range: 1,824 – 260 MCorrinne), representing an average of 14X coverage per sampleCorrinne. Two accessions, *G. thurberi* acc. 2 and G. *harknessii* acc. 7, had few retained reads and therefore were retained solely for repetitive analysis. The remaining accessions were assembled via ABySS using multiple k-mer values (see methods), and the assembly with the greatest E-size (citation) was selected for each species. E-size is an alternative assembly statistic implemented by the Genome Assembly Gold-standard Evaluations (GAGE) study to evaluate the completeness of the assembled gene space by considering the expected contig size for a randomly selected base. Here, the assembled E-sizes range from 1,066 bp in (D10) Corrinne to 10,495 bp in (D11), with an average E-size 5,356 bp (Table Assembly Stats; Figure Assembly Stats). These two species also respectively had the smallest and largest maximum contig size, i.e., 7,784 bp for (D10) Corrinne and 121,185 bp for (D11) (average = 57,076 bp Corrinne; Table Assembly Stats). Given an average gene size in the published *G. raimondii genome* of 2,583 bp, these metrics suggest that most genes will be assembled in these genomes. Indeed, in comparison to the published gene predictions for Corrinne, over 94% of genes were recovered from at least 75% of the accessions, where a gene was considered present if more than 67% of the gene was recovered from that accession. BUSCO analysis of the assembled genomes suggest general completeness of the gene space, with an average of **HOWMANY** complete BUSCOs recovered from each accession.

Corrinne: update numbers

Corrinne: update numbers

Corrinne: update numbers

Corrinne: update now that D10 has better sequence

Corrinne: update

Corrinne: update numbers

Corrinne: update numbers

Chloroplast reads were also recovered from the raw reads, representing an average of XX% Corrinne of the filtered sequencing reads. These were independently assembled (HOW). The average assembled chloroplast genome size was **HOWBIG**, which is comparable to previously published cotton chloroplast genomes (lots of citations here), and these contained an average of **HOWMANY** % ambiguity Corrinne (Table Assembly Stats).

Corrinne: include numbers

Corrinne: % N

### Gene evolution stuff

Justin's got this section/two covered. Patterns of dn/ds evolution, patterns of CNV, etc.. . .

### Phylogenetics

Right now, we have a raxml/astral tree, but idk how to get the branch lengths on there for use in ancestral state reconstructions. Probably have to make a concatenated nuclear tree for that. Could see if we could get TICR running, but it is not the easiest and may take a while to get results.. . .

### Transposable element characterization

Similar to previous reports, repetitive DNAs contribute roughly half of the total genome sequence for all species in subgenus *Houzingenia*, from an average of 44.5% in D7 to 52% in *G. anomalum D2-1*. Like most flowering plants, a vast majority of this sequence is due to the occupation of class II *gypsy* elements, which comprise

Multi-dimensional transposable element profile visualization using both log transformed and percent-genome size standardized counts showed considerable overlap among species, and even among subsections (Figure Ordination). Multivariate t-distribution confidence ellipses for each subsection overlap at least one other. 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. Selera, for example, is contained within the confidence ellipse for both Caducibractaea and Houzingenia; Integrifolia is within Houzingenia and Austroamericana. PCA visualization of the same data suggests that 31.6% and 13.8% of the variability can be explained by the first two components; however, to formerly compare the overlap in repetitive profiles 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 in Austroamericana versus both Caducibractaea 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 (**cite these**), which noted a relative abundance 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 African Ordination); that is, *G. gossypioides* is clearly lumped with the other *Houzingenia* species.

### Ancestral state reconstructions

Once the phylogeny is created, we can do ancestral state reconstructions for key TE types, criteria for selection TBD.

### Patterns of indel evolution

GATK analysis of indels, use F1 as outgroup. What sort of patterns do we want to do? Gain and loss along phylogeny, rate of gain and loss along branches? This could be complicated. Look for shared indels? Unique indels?

## Discussion

### Discussion subheading

## 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). The data were trimmed and filtered with Trimmomatic v0.32 Corrinne 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/williamssanders/D_Cottons_USDA. Corrinne

Corrinne: citation

Corrinne: Let's port this repo to a lab site after and give the new url

### Genome assembly and annotation

The trimmed data was independently assembled for each species via ABySS v2.0.1 Corrinne, using every 5th kmer value from 40 through 100. A single assembly with the highest E-size (an alternative statistic to N50; Corrinne) was selected for each species and subsequently annotated with MAKER v2.31.6 Corrinne using evidence from: (1) the NCBI *G. raimondii* EST database Corrinne, (2) *G. raimondii* reference genome predicted proteins, as hosted by CottonGen.org Corrinne, and (3) three *ab initio* gene prediction programs, i.e. Genemark v4.30 Corrinne, SNAP v2013-11-29 Corrinne, and Augustus v3.0.3 Corrinne. Both the SNAP and Augustus models were trained using BUSCO v2.0 Corrinne.

Corrinne: citation

Corrinne: citation Salzberg 2011

Corrinne: citation

Corrinne: citation

Corrinne: citation

Corrinne: citation

Corrinne: citation

Corrinne: citation

Corrinne: citation

### Gene stuff

Gene orthology and family designations were determined via OrthoFinder Corrinne...

Corrinne: citation

### Phylogenetics and ancestral state reconstruction

Trimmed reads from the genome assembly were mapped against the *G. raimondii* reference sequence (cite Paterson 2012) using BWA v0.7.10 (citation), post-processed with samtools (which version) (citation), and individual genes were independently assembled for each species/accession via BamBam v 1.4 (citation) in conjunction with the *G. raimondii* reference annotation (cite Paterson 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/williamssanders/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) (cite Ronquist and Huelsenbeck 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) (cite 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...

Measures of molecular evolution were all calculated in R v(which version)(citation). Species divergence estimates were calculated via the {chronoMPL} package (citation), using the (which time estimates?) (citation). Trees were visualized using the {ape} package (citation). 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 characterization

Reads from only one of the paired-end files (i.e., R1) were filtered and trimmed via Trimmomatic version 0.33 [?] to a uniform 95nt (https://github.com/williamssanders/D_Cottons_USDA), and then randomly subsampled to represent a 1% genome size equivalent (GSE) for each individual [?, ?]. These 1% GSEs were combined as input into the RepeatExplorer pipeline [?, ?], 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., XXX Corrinne reads from a total input of X,XXX,XXX reads) were retained for annotation as per (KOKIA CITATION), which uses the RepeatExplorer implementation of RepeatMasker [?] 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.

Corrinne: numbers please

Broad patterns of repeat occupation per genome were determined by using the abundance of each cluster in a multivariate dataset. Initial visualization of the data was conducted using both calculated in R Corrinne using both Principle Coordinate Analysis on read counts, either log normalized (to compare overall patterns of repeats) or normalized by genome size (to compare proportional cluster occupation). Differential abundance in cluster occupation was iteratively calculated in increasing phylogenetic depths to understand the evolution of repeat types throughout the evolution of the subgenus; 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 and the fitContinuous function of Geiger for visualization. All analyses are available at (https://github.com/williamssanders/D_Cottons_USDA).

Corrinne: citation

### Repeat heterogeneity and relative age

Relative cluster age was approximated using the among-read divergence profile of each cluster, as previously used for Fritillaria [?], dandelion [?], and *Kokia*/*Gossypioides*, sister outgroup genera to *Gossypium*. Briefly, cluster-by-cluster all-versus-all BLASTn [?, ?] 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 [?] to select the model with the most confidence; specific parameters can be found in (KOKIA MANUSCRIPT) and at https://github.com/williamssanders/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.

**Competing interests**
The authors declare that they have no competing interests.

**Author's contributions**
Text for this section ...

**Acknowledgements**
We would like to thank....

**Author details**
[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 39762, USA. [3]Jamie Whitten Delta States Research Center, USDA-ARS, Stoneville, MS 38776, USA.

**References**

**Figures**

**Figure 1 Sample figure title.** A short description of the figure content should go here.

**Figure 2 Sample figure title.** Figure legend text.

**Tables**

**Table 1** Sample table title. This is where the description of the table should go.

| | B1 | B2 | B3 |
|---|---|---|---|
| A1 | 0.1 | 0.2 | 0.3 |
| A2 | ... | .. | . |
| A3 | .. | . | . |

**Additional Files**
Additional file 1 — Sample additional file title
Additional file descriptions text (including details of how to view the file, if it is in a non-standard format or the file extension). This might refer to a multi-page table or a figure.

Additional file 2 — Sample additional file title
Additional file descriptions text.