







Little evidence for homoeologous gene conversion and homoeologous exchange events in *Gossypium* allopolyploids

Justin L. Conover^{1,2,3}  | Corrinne E. Grover¹  | Joel Sharbrough⁴  |
Daniel B. Sloan⁵  | Daniel G. Peterson⁶  | Jonathan F. Wendel¹ 

¹Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA 50010, USA

²Ecology and Evolutionary Biology Department, University of Arizona, Tucson, AZ 85718, USA

³Molecular and Cellular Biology Department, University of Arizona, Tucson, AZ 85718, USA

⁴Biology Department, New Mexico Institute of Mining and Technology, Socorro, NM 87801, USA

⁵Biology Department, Colorado State University, Fort Collins, CO 80521, USA

⁶Institute for Genomics, Biocomputing & Biotechnology, Mississippi State University, Mississippi State, MS 39762, USA

Correspondence

Justin L. Conover and Jonathan F. Wendel, Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA 50010 USA.

Email: jconover@arizona.edu and jfw@iastate.edu

This article is part of joint special issues of the *American Journal of Botany* and *Applications in Plant Sciences*: "Twice as Nice: New Techniques and Discoveries in Polyploid Biology."

Abstract

Premise: A complicating factor in analyzing allopolyploid genomes is the possibility of physical interactions between homoeologous chromosomes during meiosis, resulting in either crossover (homoeologous exchanges) or non-crossover products (homoeologous gene conversion). Homoeologous gene conversion was first described in cotton by comparing SNP patterns in sequences from two diploid progenitors with those from the allopolyploid subgenomes. These analyses, however, did not explicitly consider other evolutionary scenarios that may give rise to similar SNP patterns as homoeologous gene conversion, creating uncertainties about the reality of the inferred gene conversion events.

Methods: Here, we use an expanded phylogenetic sampling of high-quality genome assemblies from seven allopolyploid *Gossypium* species (all derived from the same polyploidy event), four diploid species (two closely related to each subgenome), and a diploid outgroup to derive a robust method for identifying potential genomic regions of gene conversion and homoeologous exchange.

Results: We found little evidence for homoeologous gene conversion in allopolyploid cottons, and that only two of the 40 best-supported events were shared by more than one species. We did, however, reveal a single, shared homoeologous exchange event at one end of chromosome 1, which occurred shortly after allopolyploidization but prior to divergence of the descendant species.

Conclusions: Overall, our analyses demonstrated that homoeologous gene conversion and homoeologous exchanges are uncommon in *Gossypium*, affecting between zero and 24 genes per subgenome (0.0–0.065%) across the seven species. More generally, we highlighted the potential problems of using simple four-taxon tests to investigate patterns of homoeologous gene conversion in established allopolyploids.

KEYWORDS

allopolyploid, cotton, gene conversion, homoeologous exchange, Malvaceae

Whole-genome duplications (polyploidy) are a prominent force in the evolution of plants. Polyploidy is exceptionally common in angiosperms, where it is an active ongoing process in many lineages. In addition, all angiosperms have a deep phylogenetic history that includes on average three or four rounds of polyploidy events (One Thousand Plant Transcriptomes Initiative, 2019), including one event that is shared by all flowering plants (Jiao et al., 2011). Polyploidy has also played an important role in crop

species, as many economically important crops are either currently polyploid (e.g., cotton, quinoa, potatoes) or have experienced a polyploidization event in their recent evolutionary pasts (e.g., maize, *Brassica* crops; Renny-Byfield and Wendel, 2014; Akagi et al., 2022). Understanding the dynamics of genome evolution following polyploid formation is therefore important to our understanding of plant evolution and has important potential economic and agricultural consequences.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Author(s). *American Journal of Botany* published by Wiley Periodicals LLC on behalf of Botanical Society of America.

One of the complicating factors in studying the genomes of allopolyploids (i.e., polyploids that arise through merger of divergent genomes; Wendel and Doyle, 2005; Doyle and Egan, 2010) is the possibility for physical interactions between their two (or more) co-resident genomes (i.e., subgenomes) during meiosis. Notably, the same meiotic machinery that is responsible for generating homoeologous exchanges can also lead to homoeologous gene conversion events. During the process of double-stranded DNA break repair, the broken strand of one chromatid can be repaired using its homoeologous (rather than homologous) chromosome copy. If this repair includes chiasma formation between homoeologous chromosomes (indicated by the formation of multivalents), then recombination is expected to lead to homoeologous exchanges (HEs) that reciprocally affect the region of the chromosome arm located between the chiasma and telomere (reviewed by Mason and Wendel, 2020; Deb et al., 2023). The resulting haplotype blocks of HEs are then broken up in subsequent generations via homologous recombination (and/or lost via drift), making the detection of these regions generally difficult and potentially affecting each subgenome unequally. Double-stranded DNA breaks may also be repaired via non-crossover pathways involving the homoeologous chromosome, resulting in homoeologous gene conversion (hGC, also known as nonreciprocal homoeologous recombination; Salmon et al., 2010), in which only one subgenome overwrites the other, thus distorting Mendelian segregation patterns (Lorenz and Mpaulo, 2022). Blocks of hGC are localized to the initial site of the double-stranded break and, although little is known about the typical length of hGC blocks, are substantially shorter than those resulting from HE events. Studies of homologous gene conversion in diploid *Arabidopsis* suggest that typical gene conversion blocks can range in size from tens to thousands of base pairs in length and do not appear to be biased toward creating higher GC content (Liu et al., 2018), contrary to patterns of homologous gene conversion in other eukaryotes (Taghian and Nickoloff, 1997; Lorenz and Mpaulo, 2022).

Ultimately, both HE and hGC can act to homogenize sequences across otherwise divergent subgenomes, thereby complicating allopolyploid genome assembly and analyses. In turn, this sequence homogenization can generate heterogeneous phenotypes (as demonstrated in *Brassica* [Gaeta et al., 2007; Gaeta and Pires, 2010], *Tragopogon* [Lim et al., 2008; Chester et al., 2012], *Oryza* [Li et al., 2019; Wu et al., 2021; Zhao et al., 2023]) by altering allele dosage and other (epi)genomic patterns (Bird et al., 2023), acting to reshuffle genetic variation and potentially resulting in novel genomic combinations for selection to act upon. Although multiple methods have been implemented to identify regions of allopolyploid genomes that have experienced homoeologous exchanges (e.g., competitive read mapping [Bird et al., 2021, ABBA-BABA tests [Ortiz and Sharbrough, 2023], phylogenomics [Edger et al., 2018], or chromosome staining [Chester et al., 2012]), comparatively little attention has been paid to developing methods that can identify homoeologous

gene conversion events or differentiate hGC events from HEs, particularly for older HE events that have been broken up by homologous recombination and may share many similarities with regions affected by hGC.

Homoeologous gene conversion was first described in allopolyploid cotton (Salmon et al., 2010) using expressed sequence tags (ESTs) and employing an analytical method similar to those developed to identify gene conversion in highly heterozygous diploids (most commonly created by crossing divergent, highly inbred lines) (Liu et al., 2018). In short, EST alignments were generated to include an allopolyploid cotton species (represented by both subgenomes) and its two model diploid progenitors, wherefrom homoeoSNPs (i.e., SNPs that distinguish one subgenome and its closest diploid progenitor from the other subgenome and its diploid progenitor) were identified and treated as analogous to the SNPs traditionally used in diploid-based investigations. Using this “quartet” method of comparing two homoeologous copies in a tetraploid and two orthologous copies from each of the two diploid progenitors, Salmon et al. (2010) found that hGC may affect as many as 1–2% of genes in cotton. This estimate was later updated (Flagel et al., 2012) to include additional members of the polyploid clade by using a more extensive EST data set, finding that approximately 7% of genes have been affected in one or both allopolyploid species evaluated. Subsequent efforts in evaluating hGC in cotton have relied on similar logic, albeit extending this “quartet” method to the increasingly available high-throughput sequences (Chaudhary et al., 2008; Guo et al., 2014; Li et al., 2015; Page et al., 2016) including full genomes, all of which suggest gene conversion in allopolyploid cottons is relatively rare.

Despite the clear rationale of extending these diploid individual-based “quartet” methods to an evolutionary perspective in a polyploid context, there remain a number of potential problems noted with this approach (see e.g., Salmon et al., 2010; Flagel et al., 2012; Page et al., 2016) that have not been explored. The most difficult to address is that SNP patterns indicative of hGC or HE could also arise via other evolutionary mechanisms. For example, the canonical 3:1 SNP pattern of quartet-based analyses could also be caused by (1) mutations that occur in one diploid lineage after its divergence from the actual parental lineage to the allopolyploid or (2) mutations that occur in the common ancestor of one diploid/subgenome lineage followed by another mutation at that same site in the other subgenome of the polyploid. As such, differentiating between the various evolutionary processes that can give rise to hGC/HE-like SNP patterns is important but nearly impossible without estimating the rates at which these autapomorphic (in which only one species shows a derived trait) and homoplasious (in which not all species of a monophyletic group share a derived trait) SNP patterns occur. Further exacerbating this problem, HE/hGC SNP patterns are expected to become more common in older polyploids as the evolutionary distance between the polyploid subgenomes and their diploid relatives diverge. It is important to

note that the evolutionary distance between diploids and their derivative allopolyploid subgenomes need not be equivalent, and thus the rates of autapomorphic and homoplasious SNPs need not be equal between the two lineages of the allopolyploid subgenomes. For example, if the closest extant diploid relatives to an allopolyploid differ in their relatedness to their polyploid subgenomic counterparts (i.e., if one “true” diploid progenitor goes extinct shortly following allopolyploid formation), then the terminal branch of the phylogenetic tree leading to the extant diploids would differ, and more autapomorphic SNPs would be expected to occur on the longer terminal branch. Therefore, refining the methodology to differentiate between the number of homoeoSNPs that are truly caused by hGC or HE from those caused by autapomorphic or homoplasious SNPs, as well as allowing for different rates of autapomorphic or homoplasious SNPs between the two lineages, is an important area of improvement for the analysis of HE and hGC.

Gossypium is an ideal system (Wendel and Grover, 2015; Hu et al., 2021; Viot and Wendel, 2023) to develop analytical methods for detecting hGC and HE events in allopolyploids. The genus contains ~45 currently recognized diploid species, which are classically categorized into eight genome groups (named A–G, K) based on genome size, karyotype, and patterns of intercompatibility (Endrizzi et al., 1985; Fryxell, 1992; Wendel and Grover, 2015). The genus also includes seven allopolyploids (named the AD clade), all of which are descended from the same polyploidization event (Grover et al., 2012), which occurred via hybridization 1–2 million years ago (Ma) between a member of the D lineage (most closely related to *G. raimondii* (D5)) and a member of the A genome group (equally related to *G. arboreum* [A2] and *G. herbaceum* [A1]). Economic interest in the cotton genus led to the development of high-quality genome resources for multiple species, with chromosome-scale genomes available for all seven (Chen et al., 2020; Peng et al., 2022) allopolyploids (including multiple sequences of the domesticated *G. hirsutum* and *G. barbadense*; Meng et al., 2023), 10 diploids representing the diversity of the genus, and an outgroup to the genus, *Gossypioideis kirkii*. Included within the diploid genome assemblies are the extant model diploid progenitors and their closely related diploid outgroups (Udall et al., 2019a; Grover et al., 2020), allowing for powerful analyses of post-polyploidization genome evolution. Finally, there is little chromosome number evolution within *Gossypium*, with all diploid species containing 13 haploid chromosomes (*Gossypioideis kirkii* has $n = 12$), thereby simplifying the process of developing whole-genome alignments, even in the face of a two-fold difference in genome size between the diploid lineages that gave rise to the allopolyploid (Wendel and Grover, 2015). Finally, there are hitherto no reported regions of homoeologous exchange in any *Gossypium* allotetraploid species, the presence of which could make differentiating hGCs from HEs difficult.

Here, we leverage multiple high-quality genome sequences within *Gossypium* to evaluate the extent to which we can disentangle hGC and HE from other evolutionary phenomena capable of producing similar patterns, extending the previous analyses of hGC in

Gossypium to all seven allopolyploid species. Using a well-established phylogenetic framework, we develop a robust methodology to identify potential hGCs/HE events, finding little evidence that hGC occurs in any *Gossypium* allopolyploid lineage. Nevertheless, we describe a small number of regions (~40 in total across all seven polyploids) that may have experienced hGC, HEs, or other mechanisms of inter-subgenomic sequence translocation, including the first described instance of homoeologous exchange in *Gossypium*. We discuss the implications of our work for other analyses of hGC and highlight the misleading results that may be obtained using the “quartet” method to identify hGC, especially in older allopolyploids where autapomorphic SNPs are likely and in situations where the extant diploids are not closely related to the allopolyploid subgenomes. We also discuss the role that genome assembly quality plays in the ability to identify potential hGC events.

MATERIALS AND METHODS

Data and genome alignments

Nomenclature for *Gossypium* L. (Malvaceae Juss.) species and their genomes has been standardized (Wang et al., 2018) and is used here. Specifically, the following designations are used for A-genome diploids (A1 = *G. arboreum* L.) and D-genome diploid species (D5 = *G. raimondii* Ulbr., D10 = *G. turneri* Fryxell), and F-genome diploids (F1 = *G. longicalyx* J.B. Hutch & B.J.S. Lee). In addition, the two co-resident genomes in each allopolyploid species are indicated by symbols representing their origin (A or D) along with the subscript t, for tetraploid, to distinguish them from their diploid counterparts. Genome sequences for seven allotetraploid genomes (*Gossypium hirsutum* L. [AD1; accession Bar32; Perkin et al., 2021], *G. barbadense* L. [AD2], *G. tomentosum* Nutt ex. Seem [AD3], *G. mustelinum* Miers ex G.Watt [AD4], and *G. darwinii* G. Watt [AD5], Chen et al., 2020; and *G. ekmanianum* Wittm. [AD6] and *G. stephensii* J.P. Gallagher, C.E.Grover & Wendel [AD7], Peng et al., 2022), two model diploid progenitors (*G. raimondii*, Udall et al., 2019a; and *G. arboreum*, Wang et al., 2021), an outgroup to each diploid/subgenome clade (*G. turneri*, Udall et al., 2019a; and *G. longicalyx*, Grover et al., 2020), and an outgroup to the entire genus (*Gossypioideis kirkii* (Mast.) Skovst ex J.B. Hutch, Udall et al., 2019b) were downloaded from CottonGen (Yu et al., 2021). Any scaffolds or contigs not anchored to the pseudochromosomes of each assembly were removed. Genomes for the polyploids were split by subgenome and independently aligned to a diploid reference genome using AnchorWave (Song et al., 2022) (last accessed: 6 July 2022), with annotations ported to each genome using gsnap (Wu et al., 2016). We used the proalign function within AnchorWave while allowing for the possibility of relocation variation, inversion, or chromosome fusion using flags -R 1 -Q 1 and -m 0.

Alignments of paralogous genomic regions, copy number variants, and/or presence/absence variants can easily be misinterpreted as phylogenetically discordant regions and,

hence, gene conversion or homoeologous exchange events. Therefore, we performed strict filtering to remove these regions from our whole-genome alignments. Because the size of the genomes of the diploid species of interest varies by nearly twofold (not including the fold-difference due to polyploidy per se), we took extra precautions to ensure that our analyses were not influenced by alignment errors. In particular, we aligned every genome to the smallest (*G. raimondii*, 885 Mbp) and largest (*G. arboreum*, 1700 Mbp) genomes sampled within *Gossypium* (Wendel and Grover, 2015). Pairwise alignment files were converted to gVCF files using the MAFToGVCFFPlugin tool from the Practical Haplotype Graph project (Bradbury et al., 2022), and all gVCF files were collated into a multi-sample vcf using bcftools (Narasimhan et al., 2016). Any sites including indels or non-biallelic sites were filtered out using vcftools (Danecek et al., 2011). We also ensured that, for a given diploid species or polyploid subgenome, we excluded any regions that mapped to different loci between the two diploid reference genomes using custom python scripts. Scripts for all alignments and data filtration are available on Github (<https://github.com/conJUSTover/GeneConversion>), and raw alignment and filtered VCF files are available on Figshare (DOI: 10.25422/azu.data.24512896).

Detecting homoeoSNPs and potential converted regions

To detect SNP patterns that are either indicative of potential hGC or HE events (Diagnostic + SNPs) or to create the null expected number of these SNP without invoking processes of hGC or HE (Diagnostic – SNPs), we developed a custom Python script that parses a VCF file to tabulate the total number of each SNP class, their genomic distribution, the size of consecutive Diagnostic + or – SNPs, and any SNPs that may be indicative of reciprocal hGC or HE. We developed this script to use with cases involving four taxa (i.e., two diploid progenitors and two allopolyploid subgenomes) and the full seven-taxa patterns (i.e., two polyploid subgenome, two diploid progenitors, two diploid outgroups to each subgenome/diploid clade, and an outgroup to the entire genus). This script is available in our Github repository (<https://github.com/conJUSTover/GeneConversion>). We then explored any difference in the total number of Diagnostic + and – SNPs in R and used ggplot for plotting our results.

Detecting potential regions of introgression

We used the Dsuite (Malinsky et al., 2021) package for inferences of introgression between diploid progenitors, or between polyploid species where we inferred a paraphyletic pattern of hGC. All analyses were done with window sizes of 50 SNPs, and overlapping windows of 10 SNPs.

RESULTS

An improved method to identify potential homoeologous gene conversion (HGC) events

To begin, we note that the logic presented below to describe the diagnostic SNP patterns is equivalent for hGCs as it is for HEs. Therefore, for the sake of simplicity, we will only refer to hGC SNP patterns throughout the results (except for the section below describing a shared HE event) and explore the difficulties in differentiating hGC from HE in the Discussion. Homoeologous gene conversion was initially described in allopolyploid cotton (as non-reciprocal homoeologous recombination, or NRHR) by comparing alignments of EST sequences from the two model diploid progenitor species (*G. raimondii* and *G. arboreum*) with orthologous EST sequences from the two subgenomes of allopolyploid *G. hirsutum* (Figure 1A). For the sake of generalization, we arbitrarily designate *G. raimondii* as D_1 (i.e., diploid species 1), *G. arboreum* as D_2 (i.e., diploid species 2), and the two subgenomes of an allotetraploid as P_1 (the subgenome most closely related to D_1) and P_2 (the subgenome most closely related to D_2). HomoeoSNPs were first identified as those positions where the diploid progenitors contained nucleotides each matching their respective subgenome (i.e., $D_1 = P_1$ and $D_2 = P_2$). Subsequently, diagnostic SNP patterns that may indicate HE or hGC were identified in those sites where the polyploid subgenomes were both equivalent to each other and different from one of the two parental diploids. Specifically, sites where the D_1 diploid contained a different nucleotide from that shared by D_2 , and both the P_1 and P_2 subgenomes (i.e., $D_2 = P_2 = P_1$) were considered putative HE/hGC sites in which the P_2 subgenome had “overwritten” the P_1 subgenome (Figure 1A, red box). Likewise, sites in which D_2 contained one allele, while D_1 , P_1 , and P_2 shared a second allele (i.e., $D_1 = P_1 = P_2$), were considered putative HE/hGC events where the P_1 subgenome had overwritten the P_2 subgenome (Figure 1A, blue box).

In older allopolyploids, the number of mutations that may have occurred on the terminal branches of each diploid (D_1 and D_2) following their divergence from the polyploid subgenome progenitors (P_1 and P_2 , respectively) may impact hGC diagnosis. Consider, for example, the situation where D_1 has an autapomorphy at an otherwise invariant site; this pattern mimics the diagnostic SNP pattern expected from hGC, thus potentially leading to an over-estimation of homoeologous gene conversion. Because of the possibility of these autapomorphic substitutions, the original method to detect hGC (Salmon et al., 2010) required more than one consecutive diagnostic SNP flanked by homoeoSNPs to be considered as evidence for hGC (although no flexible threshold to account for different ages of polyploids was specified). To reduce the impact of autapomorphies on inferences of hGC, we expanded the phylogenetic sampling to include species closely related to the model diploid progenitors. Specifically, we include an outgroup (D_{10} and D_{20}) for each of the diploid/subgenome

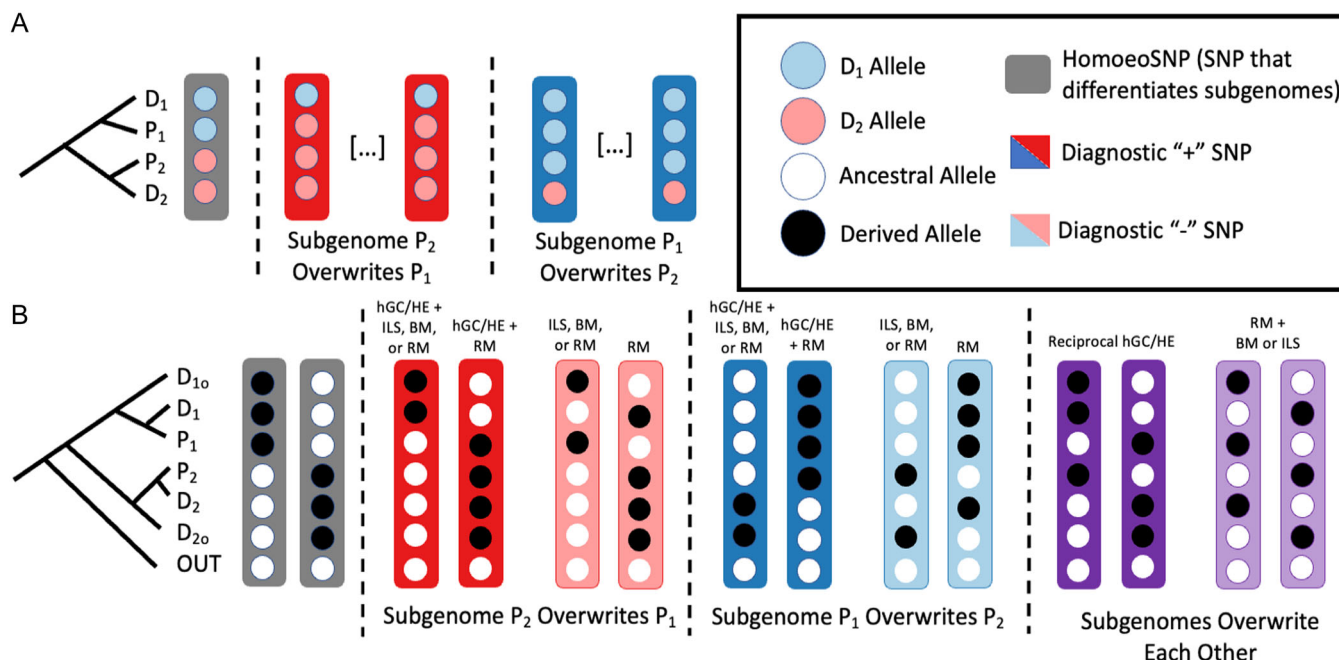


FIGURE 1 Overview of methods to identify gene conversion events. (A) Classical method of identifying hGC events. Given a four-taxon tree consisting of two model diploid progenitors (D₁ and D₂) and the two subgenomes of an allopolyploid (P₁ and P₂), categories of SNP patterns were used to infer patterns of hGC. Blue circles: SNPs in the D₁ diploid; red circles: SNPs in the D₂ diploid. First, SNPs that can reliably differentiate one diploid/subgenome clade from the other (homoeoSNPs; gray box) are identified. Between consecutive homoeoSNPs, if multiple SNPs are consistent with the pattern where subgenome P₁ has overwritten P₂ (blue box), or vice versa (red box), then gene conversion is inferred. Importantly, however, there are other evolutionary forces that may generate these SNPs patterns (e.g., autapomorphic SNPs in either diploid terminal branch) that are not accounted for in this model. (B) Our model for investigating rates of homoeologous gene conversion and homoeologous exchanges. Given a seven-taxon tree consisting of an outgroup (OUT; used to polarize SNPs as ancestral or derived), subgenomes of an allotetraploid (P₁ and P₂, respectively), two diploid progenitor species (D₁ and D₂, respectively), and an outgroup to each of the diploid/subgenome clades (D₁₀ and D₂₀, respectively), we define homoeoSNPs as those SNP sites where either the (D₁, D₁₀, P₁) or the (D₂, D₂₀, P₂) clade exclusively have derived SNPs. SNP sites that are potentially the result of hGC (Diagnostic + SNPs) are highlighted in dark blue (P₁ overwrites P₂) or dark red (P₂ overwrites P₁). For each of these SNP patterns, multiple other mutational/evolutionary patterns may result in the same pattern, including recurrent mutation (i.e., mutation on internal branch leading to D₁/D₁₀/P₁ and a separate mutation on the terminal branch leading to P₂), back mutation (i.e., mutation on the internal branch leading to D₁/D₁₀/P₁ followed by a separate mutation on the terminal branch of P₁ to revert back to the ancestral state), or incomplete lineage sorting (i.e., when a mutation occurs shortly before the divergence of D₁₀ from the D₁/P₁ clade, but sorts into a paraphyletic pattern). For each Diagnostic + SNP pattern, we compare the genome-wide frequency to the SNP patterns that exclude gene conversion (Diagnostic - SNPs; light blue, light red), assuming that these should occur at equal frequencies in the absence of gene conversion. That is, any hGC events will increase the number of SNP patterns consistent with hGC, while SNP patterns consistent with ILS, back mutation, or recurrent mutation should remain unaffected. Therefore, we expect the ratio of these SNP categories to be equal if no gene conversion is present, whereas a higher number of gene conversion than Diagnostic - SNP patterns would indicate hGC or homoeologous exchange.

clades to phylogenetically diagnose autapomorphic SNPs in each diploid (D₁ and D₂), thereby allowing the identification and removal of these putative hGC sites as potential sources of error. Additionally, we included a phylogenetic outgroup to the entire genus to determine the internal branch on which these hGC-informative SNPs arose, allowing us to consider both bias in the direction of gene conversion (toward P₁ or P₂) and the extent to which derived alleles revert (or convert) back to their ancestral state.

This expanded phylogenetic sampling produced a set of putatively diagnostic SNPs that only includes sites where the diploid progenitor (e.g., D₁) and its outgroup (e.g., D₁₀) harbor derived alleles or are the only two with ancestral alleles (Figure 1B, dark red boxes). Quantification of these SNP patterns, however, does not directly measure the rate of hGC as other evolutionary phenomena could also give rise to these SNP patterns. For example, the SNP patterns in

which only D₁ and D₁₀ contain derived alleles (Figure 1B, first red box from left) could also be caused by three additional mechanisms: (1) separate, recurrent mutations on the terminal branches of D₁ and D₁₀; (2) a mutation in the common ancestor of D₁, D₁₀, and P₁ that undergoes incomplete lineage sorting (ILS) to transmit the derived allele to only D₁ and D₁₀; and (3) a mutation in the common ancestor of D₁, D₁₀, and P₁, followed by a back mutation in the P₁ terminal branch subsequent to allopolyploid formation.

While these additional evolutionary scenarios may initially appear to unnecessarily complicate inferences of hGC, we can leverage the symmetrical properties of a phylogeny to estimate the frequency with which these scenarios affect our diagnostic SNPs. Since we expect these other evolutionary phenomena (e.g., ILS) to be distributed across branches independently of ploidy level, we can

assume they produce hGC-like phenomena equally across symmetrical branches, forming a baseline estimate for each. For example, in scenario 1 (recurrent mutation), we can assume that the mutation rate on the terminal branch of D_1 is equal to that on the terminal branch of P_1 . Thus, if we find an equal number of SNP patterns that can be explained by recurrent mutation on the D_1 and P_1 terminal branches (when there is also a mutation on the terminal branch of D_{10} , as explained above), respectively, we can infer that there is no hGC present in our samples. Homoeologous gene conversion would therefore be indicated by an excess of SNP patterns consistent with hGC (henceforth, “Diagnostic + SNPs”) compared to SNP patterns that can only be explained by recurrent mutation (henceforth, “Diagnostic – SNPs”). Similarly, under ILS (scenario 2), there is an equal probability that the derived alleles will be present in D_{10}/D_1 versus D_{10}/P_1 (but not D_1/P_1 due to more recent shared phylogenetic history); therefore, in the absence of hGC, we would expect the number of hGC-like SNP patterns (Diagnostic + SNPs) to equal those explained by the other ILS patterns (Diagnostic – SNPs). Finally, to estimate the influence of back mutations (scenario 3), we expect that the number of back mutations in the P_1 terminal branch is equal to the number of back mutations in the D_1 terminal branch. Thus, if no hGC has occurred, we would expect an equal number of Diagnostic + SNP patterns indicative of gene conversion as those where only D_{10} and P_1 contain the derived allele (i.e., Diagnostic – SNPs).

The second category of SNPs that may indicate hGC includes those in which both D_1 and D_{10} are the only species in the phylogeny that contain the ancestral allele (Figure 1B, second red box from left). These SNP patterns may arise via only two evolutionary scenarios: recurrent mutation and hGC. Recurrent mutation would occur by a mutation in the common ancestor of $D_2/D_{20}/P_2$, followed by a recurrent mutation in the terminal branch of P_1 . Because we expect the number of mutations occurring on the terminal branch of P_1 to be the same as those occurring on the terminal branch of D_1 , the number of SNP patterns caused by recurrent mutations in which only D_1 and D_{10} contain the ancestral allele (i.e., Diagnostic + SNPs) should be equal to the number of SNP patterns in which only D_{10} and P_1 contain the ancestral allele (i.e., Diagnostic – SNPs). Thus, any excess in the number of Diagnostic + SNPs relative to Diagnostic – SNPs would be evidence for hGC.

The above logic and scenarios are applicable and symmetrical with respect to the direction of hGC (in the present case where subgenome P_1 overwrites P_2 [Figure 1B, dark blue boxes]) and can be extended to situations in which the subgenomes have reciprocally “overwritten” each other (Figure 1B, purple boxes) either through homoeologous exchange(s) or if multiple hGC events affect the same locus (but in opposite directions) in different individuals and/or at different timepoints. Therefore, our test for hGC not only considers the presence of SNP patterns consistent with hGC (Figure 1B, dark red and dark blue boxes), but also evaluates the abundance of these Diagnostic + SNPs

relative to SNP patterns that can be explained by other evolutionary mechanisms (Figure 1B, light red and light blue boxes), providing a baseline measure for these confounding phenomena. As such, the presence of hGC can be inferred by an excess of Diagnostic + SNPs, with the absence of hGC being evidenced by equal numbers of Diagnostic + and – SNPs in our data set. Furthermore, this method may also be useful in identifying regions of the genome that have experienced other mechanisms of reciprocal homoeologous recombination (Figure 1B, dark purple boxes), including homoeologous exchanges, which we expect to affect larger regions of the genome and to be biased toward the distal regions away from the centromeres. These reciprocal homoeologous recombination patterns are notoriously difficult to identify, however, because they result in no change in allelic dosage and because they may be artificially created by genome assembly errors, in part due to incorrect subgenome assignment.

While the aforementioned logic can be applied to genome-wide SNP counts, it is also important to consider approaches to identify particular genomic regions that have experienced hGC or HE and the direction of these exchanges. We can use unaffected homoeoSNPs as potential “outer bounds” for regions in the genome containing Diagnostic + SNP patterns where a potential hGC could have occurred (unless the gene conversion event resulted in a mosaic of converted and unconverted homoeoSNPs, which we discuss below). For example, by comparing the number of regions in which there are three sequential Diagnostic + SNPs versus three sequential Diagnostic – SNPs in the same direction, we can calculate the proportion of those regions that have experienced hGC using a simple D statistic. In cases where homoeologous gene conversion leads to a mixture of converted and unconverted homoeoSNPs (which would occur under GC-biased gene conversion, for example), we would still expect an excess of Diagnostic + SNPs, but would expect that each hGC tract would lead to multiple regions indicative of hGC; however, these regions would be expected to harbor fewer Diagnostic + SNPs and be smaller in size on average compared to homoeologous gene conversion tracts in which all nucleotides from one subgenome are converted.

A final consideration in the methodological logic concerns the possibility of interspecific gene flow, which can create SNP patterns similar to that of hGC, thereby complicating its detection. There are several scenarios of hybridization that may lead to similar SNP patterns as those indicative of hGC. For example, a mutation that occurs in the D_{20} terminal branch, followed by gene flow from D_{20} into the ancestor of D_1 , D_{10} , and P_1 would create SNP patterns in which mutations are shared by D_{10} , D_2 , D_{20} , and P_2 (which is a Diagnostic + SNP pattern). Likewise, mutations that occur in the common ancestor of D_2 , D_{20} , and P_2 , followed by gene flow from any of these lineages into D_1 , would create a Diagnostic – SNP pattern, thereby leading to an underestimation of the rate of hGC. Additionally, introgression between different ploidy levels

(i.e., interploidy introgression) can influence these relative rates of Diagnostic SNPs in similar ways. Although allopolyploidy is considered a strong mechanism of reproductive isolation from diploid progenitors, recent studies suggest that interploidy introgression may be more common than previously realized (Kryvokhyzha et al., 2019; Wang et al., 2023). Therefore, care should be taken when using these methods in systems with likely or recurrent gene flow, removing those genomic regions influenced by hybridization before interpreting results of hGC.

No history of interspecific hybridization in *Gossypium* diploids or interploidy introgression

Because detection of hGC using a phylogenetic SNP-based approach may be biased in the presence of introgression between diploid groups or by interploidy introgression, we sought to identify potentially introgressed regions using the analytical framework of the Dsuite package (Malinsky et al., 2021) across the entire phylogeny used here, including all seven allopolyploids. Although we did find trios that suggest widespread hybridization amongst the polyploids, none of these trios involved inter-subgenomic hybridization as would be expected under homoeologous exchanges, and we did not find strong evidence of introgression in any interploidy comparisons or comparisons involving only diploid species (Appendix S1: Figure S1; Appendix S2). Unsurprisingly, we found no evidence of any introgression in any of the trios tested, presumably because our two diploid clades are from Central America and Africa and have remained separated by the Atlantic Ocean since their divergence 5–10 million years ago (Ma), aside from the apparently ephemeral contact 1–2 Ma that resulted in the polyploid clade (Wendel and Grover, 2015). This analysis, however, is necessary as there remains the possibility, however small, that following the migration of an A-genome propagule to the American continents, there could have been introgression from the A-genome group into the progenitor species of the polyploid D-subgenome. Evidence for introgression of rDNA and other repeated sequences has been previously suggested in *G. gossypoides* (Wendel et al., 1995; Zhao et al., 1998; Cronn et al., 2003), although genomewide analyses have not replicated this finding for the rest of the nuclear genome or for additional species (Grover et al., 2019), suggesting that this introgression may have been limited to *G. gossypoides*.

Identifying potential regions of homoeologous gene conversion in *Gossypium*

Because all seven polyploids in *Gossypium* diverged from the same polyploidization event (Grover et al., 2012; Hu et al., 2021), we used the same set of diploids to identify our Diagnostic +/- SNP patterns in each polyploid species.

Namely, the A-diploid progenitor (D₁, Figure 1B) is represented by *G. arboreum* (species label A2; see Materials and Methods for genome designations); the A-diploid outgroup (D_{1o}, Figure 1B) is represented by *G. longicalyx* (species labeled F1); the D-diploid progenitor (D₂, Figure 1B) is represented by *G. raimondii* (species label D5); and the D-diploid outgroup (D_{2o}, Figure 1B) is represented by *G. turneri* (species label D_{1o}). All SNPs were polarized into ancestral or derived states using *Gossypoides kirkii*, an outgroup that diverged from *Gossypium* circa 6–12 Ma (Udall et al., 2019b). We initially analyzed each polyploid independently, treating the A-subgenome (i.e., “At” for “A-tetraploid”) and the D-subgenome (i.e., “Dt”) as P₁ and P₂, respectively (Figure 1B).

For each polyploid, we were able to identify over 1 million homoeoSNPs that distinguish the two homoeologous subgenomes (ranging from 1.01 million in AD2 to 1.05 million in AD1; Appendix S1: Figure S2). While the number of homoeoSNPs is considerably lower than previously reported (~25 million; Page et al., 2013), we note our strict requirements for homoeoSNP definition. That is, homoeoSNPs were only inferred when all members of the D5/D10/Dt clade (where “Dt” refers to the D-subgenome in the tetraploid) shared a SNP that was different from the SNP shared by all members of the F1/A2/At clade (Figure 1B, grey box). As expected, more derived alleles were found at the base of the D clade compared to the A clade (Appendix S1: Figure S2) due to the relatively more recent divergence of D5/D10/Dt (~1.76 Ma; Grover et al., 2019) versus F1/A2/At (~4 Ma; Grover et al., 2020). Additionally, there were more derived mutations present on the terminal branches of both subgenomes of all polyploids (with the exception of the At subgenome of *G. mustelinum* [AD4]) compared to their respective diploid progenitor (Appendix S1: Figure S2), consistent with previous findings (Chen et al., 2020) and which may be a result of the masking effects of allopolyploidy that reduces the fitness consequences of deleterious alleles (Conover and Wendel, 2022) or from unequal evolutionary rates between the subgenomes as compared to their diploid relatives (Sharbrough et al., 2022).

Homoeologous gene conversion where Dt overwrites At

For each polyploid, we identified between 24,900 and 26,000 regions that were flanked by homoeoSNPs and also contained at least one Diagnostic +/- SNPs (Figure 2A) that would be consistent with the Dt subgenome overwriting the At subgenome. These regions ranged in size from a single nucleotide (i.e., two homoeoSNP had a single base pair between them, and that base pair showed a SNP pattern consistent with a +/- Diagnostic SNP) to over 576 kb, with a strong bias toward smaller regions. Only a single Diagnostic +/- SNP was identified in most regions (Figure 2A), and the number of identified regions decreased

as the number of consecutive Diagnostic +/- SNPs increased. The total distribution of Diagnostic + SNPs compared to the distribution of Diagnostic - SNPs showed no statistical significance ($P > 0.99$ for all species, two-sided Kolmogorov-Smirnov test); thus, for graphical clarity, we combined all regions in which the number of Diagnostic +/- SNPs in all seven species ranged from 5 to 11 (5 being the smallest bin size in which some species had fewer than 10 regions, and 11 being the largest number of consecutive Diagnostic - SNPs, although these are arbitrary cutoff points used only for graphical clarity). The region with the largest number of Diagnostic - SNPs (i.e., those consistent with ILS or recurrent mutation, but not hGC or HE) in all species contained 11 SNPs, while the region with the highest number of Diagnostic + SNPs contained 150 SNPs. Interestingly, these higher numbers of Diagnostic + regions were clustered at the terminus of chromosome D5_01 in six of the seven polyploids in our analysis, indicating a

homoeologous exchange event rather than a hGC event (discussed below).

Because the Diagnostic + SNP patterns can reflect evolutionary processes other than homoeologous gene conversion, we compared the proportion of Diagnostic + to Diagnostic - SNPs in each polyploid species to assess the putative rate of hGC. If hGC has historically occurred in any of these lineages, we expect to see a marked excess in the proportion of Diagnostic + SNPs relative to Diagnostic - SNPs. In contrast to our a priori expectations based on earlier hGC assessments, we saw no enrichment of Diagnostic + SNP patterns (relative to Diagnostic - SNPs) for those regions containing four or fewer consecutive diagnostic SNPs, either when comparing the total number of SNPs within regions (Figure 2B) or in the proportion of SNPs that exhibit Diagnostic + SNP patterns compared to the total number of SNPs (Figure 2C). Interestingly, we saw a higher fraction of regions that contain Diagnostic - compared to

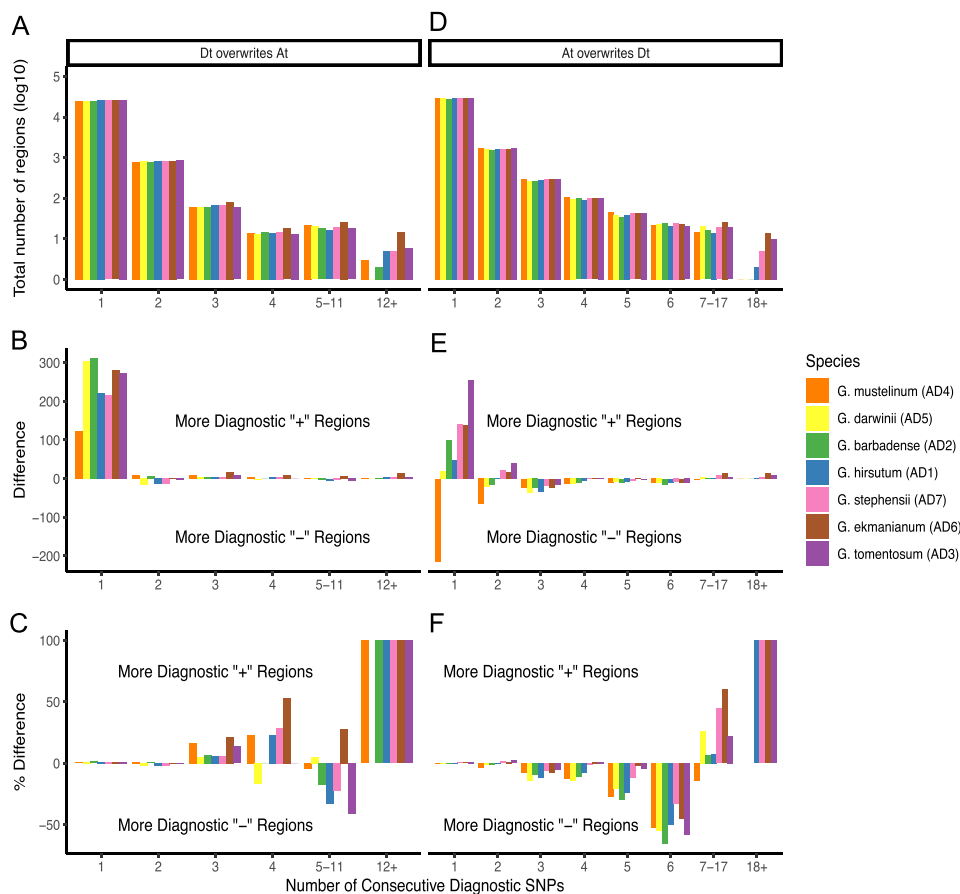


FIGURE 2 Patterns of SNPs indicative of homoeologous gene conversion and homoeologous exchange. Regions of the genome indicative of the Dt subgenome overwriting the At subgenome (A, B, and C) or the reciprocal direction (D, E, and F) were identified by first identifying homoeoSNPs (i.e., SNPs in which a subgenome, its most closely related diploid progenitor, and diploid outgroup all have one allele, while the other subgenome and its most closely related diploid progenitor and diploid outgroup have a different allele at that site). The number of Diagnostic + or Diagnostic - SNPs located between two flanking homoeoSNPs is tabulated along the x-axis of each plot. (A, D) Total number of regions (y-axis) with a given number of consecutive Diagnostic + and - SNPs (x-axis). (B, E) The difference in the number of regions with Diagnostic + compared to Diagnostic - SNPs. Positive values indicate more regions with Diagnostic + SNPs, negative values indicate more regions with Diagnostic - SNPs. (C, F) The percentage difference between the number of regions with Diagnostic + or - SNPs, calculated as (difference between Diagnostic + and Diagnostic - regions/total number of regions), where positive values indicate more regions with Diagnostic + SNPs. Each polyploid is represented by a different color in the bar graph, shown at right.

Diagnostic + SNPs in several of these categories, potentially indicating a higher rate of recurrent mutations on the terminal branches of the polyploids as compared to the terminal branches of the diploid progenitors. For regions that only have a single Diagnostic SNP (Figure 2B), we see a higher number of Diagnostic + SNPs for all species, indicating that a small amount of hGC may be occurring that predominantly affects small regions of the genome (and/or that these gene conversion events may have led to a mix of converted and unconverted SNPs, artificially lowering the size of these regions); however, because we only see 100–300 excess regions out of a possible 10,000 total regions, this difference is not statistically significant and should not be interpreted as evidence for hGC taking place.

Homoeologous gene conversion where At overwrites Dt

For those SNP patterns that are consistent with hGC in the opposite direction, (i.e., At overwriting Dt; Figure 2D), we see a slightly higher number of regions (29,000–32,000), as described above for potential hGC, presumably due to a more distantly related diploid outgroup on this side of the phylogeny (*G. longicalyx*) compared to the other (*G. turneri*), allowing for a larger proportion of sites with SNP patterns caused by recurrent mutation to pass filtering. The broad size patterns of these SNPs across the genome are similar to those described above (ranging in size from 1 bp to 622 kb), where regions in which one Diagnostic + or – SNP is contained within a single region flanked by homoeoSNPs are the most common. The region with the highest number of Diagnostic – SNPs contained 11 SNPs, and the total distribution of Diagnostic + compared to the distribution of Diagnostic – SNPs showed no statistical significance ($P > 0.88$ for all species, two-sided Kolmogorov–Smirnov test). Therefore, we combined some groups for graphical clarity depending on the total number of sites present in each species; namely, all species had at least 10 sites with six or fewer consecutive Diagnostic + SNPs, and the highest number of consecutive Diagnostic – SNPs in any species was 17, so we combined groups of size 7–17. The difference in the highest number of consecutive Diagnostic + to Diagnostic – SNPs is different for the two directions of potential hGC or HE (17 when Dt overwrites At, 11 when At overwrites Dt) due to the difference in the phylogenetic relatedness of the diploid outgroups in the two clades and is an important aspect of choosing samples for repeating this analysis in other systems (see Discussion). The regions with the most Diagnostic + SNPs in any species contained 105 (chromosome D5_05 in *G. tomentosum* [AD3]) and 76 (chromosome D5_12 in *G. ekmanianum* [AD6]) such SNPs; however, because these regions are restricted to a single species and do not occur at the termini of the chromosome arms, as would be expected under HE events, it is difficult to differentiate whether these patterns are caused by real hGC events or are due to artifacts such as

alignment errors or genome assembly errors. Notably, *G. ekmanianum* and *G. tomentosum* contained considerably higher numbers of regions with elevated numbers of Diagnostic + SNPs (Figures 2F, 3), indicating that these genome assemblies (or alignments) are of poorer quality than those of the other polyploids in the clade or that there may be a temporal and lineage-specific variation in the rate and occurrence of HE and/or hGC in *Gossypium*. When comparing the difference in the number of regions with Diagnostic + or Diagnostic – SNPs indicative of At overwriting Dt (Figure 2E), we again see similar patterns as described above (Figure 2B). Congruent with our prior analysis, we saw no evidence of enrichment in Diagnostic + SNPs (versus Diagnostic – SNPs) for regions of size bins 1 or 2, although in most species, a higher number of regions have a single Diagnostic + SNP compared to regions that have a single Diagnostic – SNP, with the exception of *G. mustelinum*. As mentioned above, because this small difference is tabulated from a total of over 10,000 regions, this result is not statistically significant and should not be interpreted as evidence of hGC taking place.

Potential gene conversion or homoeologous exchange events where Diagnostic + SNP tracts are longer than Diagnostic – SNP tracts

In six of the seven allopolyploids (i.e., all except *G. darwinii*, AD5), there was at least one genomic region that contained more consecutive Diagnostic + SNPs than the longest track of consecutive Diagnostic – SNPs observed in any species. In total, we found 19 regions of this type that are consistent with the At subgenome overwriting the Dt subgenome (Appendix S1: Figure S4). Two of these regions were found in more than one species (with identical locations for flanking homoeoSNPs) and occur in parsimonious positions along the polyploid phylogeny (Figure 3). One region found on chromosome D5_02 in *G. hirsutum* (AD1), *G. tomentosum*, (AD3), *G. ekmanianum* (AD6), and *G. stephensii* (AD7) is 3888 base pairs long, contains twenty consecutive Diagnostic + SNPs, and partially overlaps with two genes. Because these four polyploids form a monophyletic group, it is more likely that this region reflects an event in the common ancestor of these four species following their divergence from the other three polyploid species (*G. mustelinum* (AD4), *G. barbadense* (AD2), and *G. darwinii* (AD5)) rather than an error in genome alignment. A second region on chromosome D5_13 was 2458 base pairs in length and partially overlaps one gene. This region, however, was found in *G. tomentosum* (AD3), *G. ekmanianum* (AD6), and *G. stephensii* (AD7), which form a monophyletic group only with the inclusion of *G. hirsutum* (AD1). Interestingly, however, we found evidence that this region in the *G. hirsutum* genome may have experienced introgression from *G. barbadense* (AD2) (Appendix S1: Figure S3), either as a result of historical natural introgression between these two species or as a result of the known intentional introgression during crop domestication and

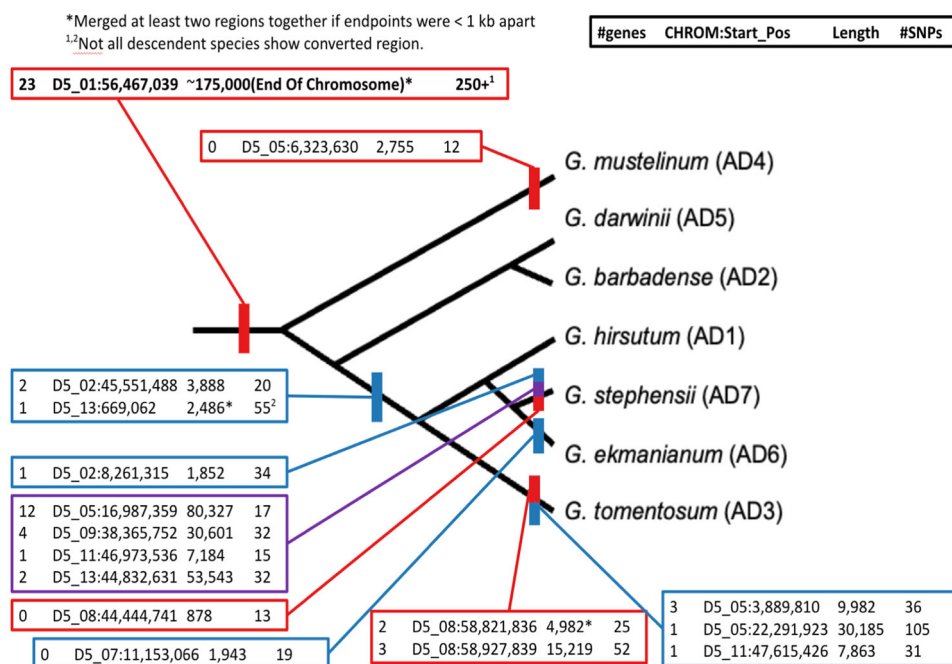


FIGURE 3 Evolutionary timing of potential homoeologous gene conversion or homoeologous exchanges in *Gossypium* phylogeny. Using only the regions that were longer than the longest region consistent with patterns of Diagnostic – SNPs (i.e., at least 12 SNPs when Dt overwrites At [outlined in red], and at least 18 SNPs when At overwrites Dt [outlined in blue], and at least 4 SNPs where reciprocal hGC has occurred [outlined in purple]), we can map each of these events to the *Gossypium* phylogeny using parsimony. For each line, the four columns represent the number of genes in the region, the chromosome/starting point of the region, the length of the region, and the number of consecutive Diagnostic + SNPs in that region, respectively. If any region included more than three SNPs indicative of reciprocal hGC, then the text of that region is purple, and the number of reciprocal SNPs is included in parentheses. Any regions that were subsequently attributed to errors in genome assembly in *G. ekmanianum* (AD6) and *G. tomentosum* (AD3) have been removed, but are still detailed in Appendix 1 (Figure S4). The phylogeny does not represent scaled evolutionary distances or divergence between species, only the relative relationships as inferred from previous analyses (Chen et al., 2020; Peng et al., 2022).

improvement (Yuan et al., 2021). Each of the remaining 17 regions that contained long stretches of Diagnostic + SNPs (affecting 0–6 genes each, for 32 total) were confined to a single species, potentially revealing recent hGC and/or HE events, although errors in genome assembly or alignments are likely the cause of many of these regions, which we explore below.

Similarly, we also found 16 regions consistent with the Dt subgenome overwriting the At subgenome in which the number of Diagnostic + SNPs was higher than any region with consecutive Diagnostic – SNPs, and only one of these regions was shared by more than one species. This single region was shared by all species except *G. darwinii* (AD5) and affected the terminal 175 kb (23 genes) of chromosome D5_01. The number of putatively converted SNPs in this region varied from 120 to 261 among species (see below), but the “starting point” (i.e., the homoeoSNP most distant from the chromosome terminus) was consistent in all six polyploids, resulting in the complete conversion of 23 genes. We explore this region in detail below. The remaining 15 regions (affecting 0–8 genes each) were all present in a single species, again with the majority found in *G. ekmanianum* (10 regions, 24 genes) and the remainder found in *G. tomentosum* (3 regions, 6 genes), *G. stephensii* (1 region, 0 genes), and *G. mustelinum* (1 region, 0 genes).

For those regions that were localized to only *G. ekmanianum* (AD6; 20 regions) or *G. tomentosum* (AD3; 9 regions), we investigated whether these patterns could be caused by genome assembly errors at or near these regions. By mapping the original CLR read used in genome assembly back to the assembled genome, we visualized the mapping quality and read depth in each subgenome to look for patterns consistent with errors in genome assembly, including low read depth, considerable inconsistencies (i.e., sharp changes) in read depths, or where a portion of the assembly was supported by zero reads (e.g., an insertion was included in the final assembly that is not supported by any of the CLR reads). We found that 19 of the 20 of the regions that were restricted to *G. ekmanianum* (Appendix S1: Figure S5; Appendix S3: Table S1) showed clear signs of assembly error (although the flanking region around the 20th regions showed potential signals of assembly errors), and that four of the nine regions restricted to *G. tomentosum* showed clear assembly artifacts (Appendix S1: Figure S6; Appendix S3: Table S2). We also investigated the three regions that were shared by *G. ekmanianum* and *G. tomentosum* (and various other species, depending on the region) and found no obvious signs of assembly artifacts in any of these regions. Thus, while homoeologous exchanges and homoeologous gene conversion can complicate allopolyploid genome assembly, we also

highlight that complications in allopolyploid genome assembly can also lead to false inferences of homoeologous exchanges and homoeologous gene conversion.

A single, shared homoeologous exchange event occurred shortly after polyploidization

The region described above, involving the final 175 kbp and 120–261 SNPs on the end of chromosome D5_01, is indicative of a HE event that results from recombination between the subgenomes of an allopolyploid in which the Dt subgenome has overwritten the At subgenome. Typically, HE events affect the entire terminus of a chromosome past the recombination breakpoint, and this region is then fragmented by homologous recombination in subsequent generations in an analogous way that blocks of introgression are broken up by recombination. This region, however, contains few SNPs that are indicative of recombination breaking up the region in any of the polyploids, suggesting that the initial recombination event leading to the HE either occurred in a very small population, perhaps immediately following the bottleneck that is typically associated with allopolyploid formation, or has experienced strong positive selection, leading to the fixation of the HE in the ancestral population before the first speciation event occurred. Supporting this hypothesis, we note that in the six polyploids that show evidence of this HE event, all displayed the same initial recombination point, and all contained few places of homologous recombination that broke up this HE event into smaller regions. Notably, one species in our analysis (*G. darwinii*, AD5) failed to show evidence of this HE event that is shared by all other polyploid species; however, further inspection suggests this lack of evidence may be an assembly artifact (Figure 4). Inspection of genome alignments between all of the polyploids and their diploid progenitors revealed a gap in the assembly at the ends of both chromosomes D01 and A01 in *G. darwinii*, suggesting that the initial lack of evidence for this HE may be the result of inter-subgenomic sequence similarity (due to the HE) impeding the accurate assembly of homoeologous chromosomes A01 and D01 in *G. darwinii*. Interestingly, this region also exhibits missing sequences in some of the other allopolyploid genomes (e.g., *G. barbadense* [AD2]). These gaps, however, did not span the entire HE region, and thus did not inhibit our ability to detect the HE event (Figure 4B).

Reciprocal hGC and HE

Finally, we evaluated SNP patterns consistent with reciprocal gene conversion where we simultaneously observed consecutive At SNPs on the Dt chromosome and vice versa for the same region (Figure 1B, dark purple boxes). Although the mechanism of hGC is inherently unidirectional, it is possible that hGC may occur in different

individuals in different directions at the same time, thus leading to hGC in both directions. Additionally, these SNP patterns may be caused by HEs (which are inherently reciprocal) or by errors in genome assembly where the subgenome assignment is incorrect. While it is difficult to estimate the expected frequencies of counterbalancing SNP patterns that could be explained equally well by recurrent mutations, ILS, and/or back mutations (as we created for the previous gene conversion analyses), we found these SNP patterns to be rare, indicating that reciprocal hGC or HE is not likely to have occurred in any *Gossypium* allopolyploid. In four of the polyploids (*G. hirsutum* [AD1], *G. barbadense* [AD2], *G. mustelinum* [AD4], and *G. darwinii* [AD5]), we found three or fewer regions with two consecutive SNPs indicative of reciprocal gene conversion (and never saw more than two consecutive SNPs fitting this pattern). We observed longer regions with putative reciprocal gene conversion SNP patterns in the other three polyploids (i.e., *G. tomentosum* [AD3], *G. ekmanianum* [AD6], and *G. stephensii* [AD7]), although we approach these with caution. In *G. tomentosum*, we identified a single region that contained 57 consecutive SNP patterns consistent with reciprocal gene conversion and a second region that contained only two neighboring SNPs congruent with reciprocal hGC. We note, however, that *G. tomentosum* was among those species mentioned above where assembly errors artifactually generated observations of hGC and that these assembly errors could produce SNP patterns consistent with reciprocal hGC. Likewise, both *G. ekmanianum* (19 regions, 3–78 SNPs) and *G. stephensii* (4 regions, 15–32 SNPs) exhibited SNP patterns consistent with reciprocal gene conversion, and while these regions were among the largest discovered in our analysis (between 7 KB and 80 KB, affecting 1–12 genes each), we note that these genomes were also among those suspected of assembly artifacts, similar to *G. tomentosum*.

A direct comparison of quartet-based methods

While the primary goal of this study is to develop an updated method to detect hGC or HE events while explicitly accounting for other evolutionary processes that may create similar SNP patterns (e.g., ILS, recurrent mutation), we also tested the classical “quartet” approach on a genomewide basis by tabulating the number of Diagnostic + SNPs (Figure 1A; i.e., SNP patterns that may be created via hGC or autapomorphic mutation on a diploid terminal branch) as well as SNPs that are analogous to our Diagnostic – SNPs in the seven-taxon test (i.e., SNP patterns that may be created via autapomorphic mutation on a terminal polyploid subgenome branch). Given no instances of hGC or HE, we expect that these distributions should be identical, and any deviations in which Diagnostic + SNP patterns are observed more frequently than Diagnostic – SNP patterns would suggest that hGC or HE may play a role in shaping the overall SNP patterns in the genome.

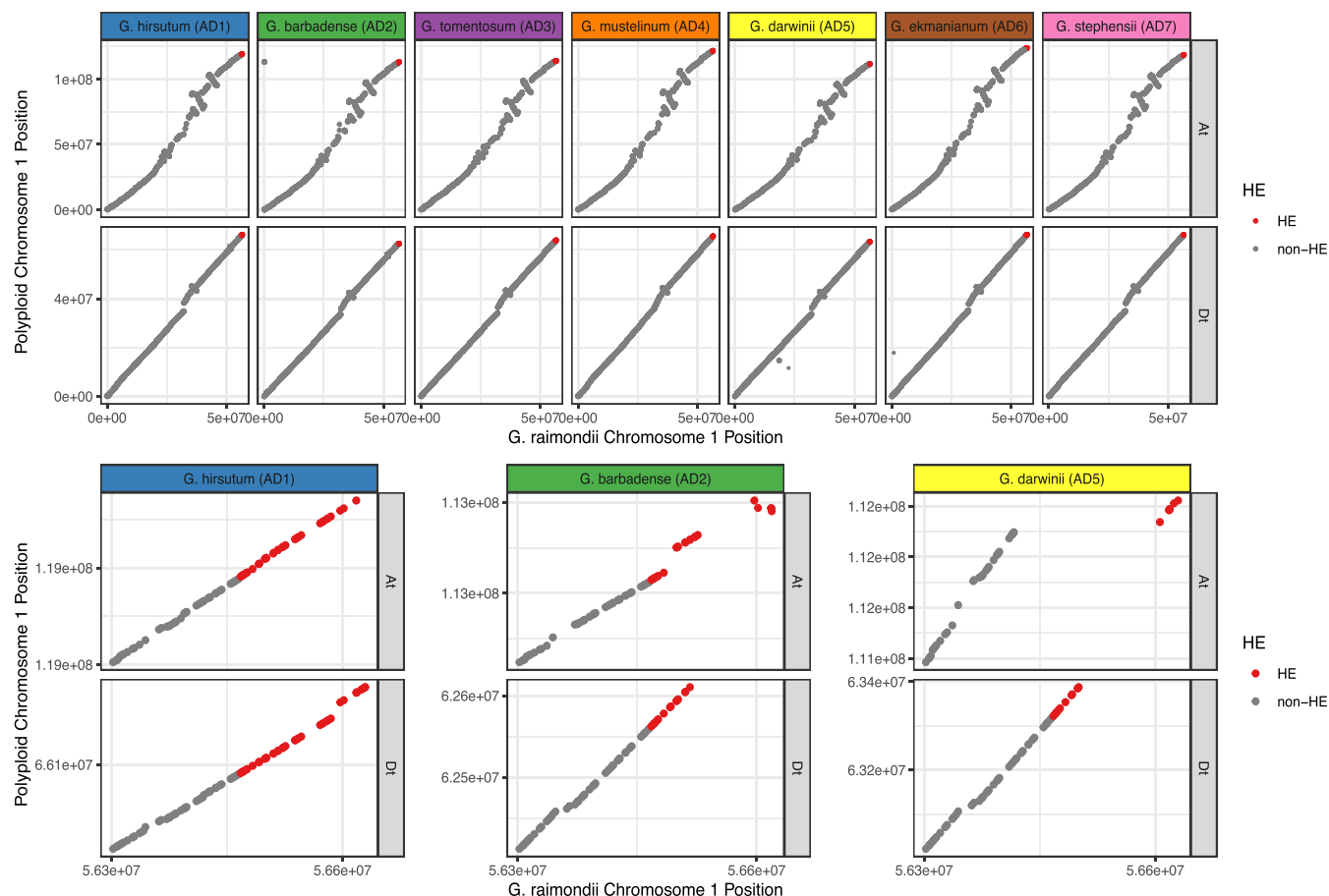


FIGURE 4 Homoeologous exchange event on chromosome 1 locates to a gapped region in all polyploid genome assemblies. (A) Dot plot showing the whole-genome alignment between each subgenome (top: At subgenome; bottom: Dt subgenome) aligned to the *G. raimondii* genome. Anchors identified by Anchorwave are shown here. Colors of the facet labels are consistent with the color scheme used through the other figures. (B) Dot plot of the region that has experienced a homoeologous exchange event. Regions that are inferred to have experienced the HE event are in red, and regions that did not experience the HE event are in grey. In *G. hirsutum* (left, blue), accurate assembly of both subgenomes allowed for the accurate identification of the HE event. In *G. barbadense* (middle, green), a gap in the At subgenome (top) and incomplete assembly of the telomeric region of the Dt subgenome (bottom) did not hinder HE identification, even though the HE is the likely cause of the assembly difficulties, as the regions directly flanking the initial recombination event (i.e., where the grey dots meet the red dots) is still intact in both subgenomes. In *G. darwinii* (right, yellow), a large gap in the At subgenome (top) and incomplete assembly of the telomeric region of the Dt subgenome (bottom) led to the inability to positively identify this region as an HE since there is no vertical region in which there are two anchors shown in red.

Using the quartet approach, we found nearly three orders of magnitude more regions with SNP patterns that fall into our Diagnostic + and - categories (2.14 M [AD2] to 2.22 M [AD1]) relative to our 7-taxon test (52.2 K [AD2] to 54.7 K [AD1]), likely due to the less stringent SNP thresholds in the quartet test. Thus, a higher number of these SNPs are likely due to autapomorphic mutations, recurrent mutations, or back mutations. The distribution of consecutive Diagnostic + SNPs was much broader than our 7-taxon test, with the highest number of consecutive Diagnostic + SNPs being 347 (the HE region on chromosome 1 in AD4) and the highest number of consecutive Diagnostic - SNPs between flanking homoeoSNPs being 2140 (chromosome 12, AD5). For hGC or HE occurring in either direction, we found that regions with only a single Diagnostic + or - SNP occurred most frequently (Figure 5A, B) and that the number of regions decreased as the number of consecutive SNP patterns increased. Additionally, we found that the distribution became increasingly dissimilar between

species as the total number of consecutive SNPs increases, indicating that the upper tail of these distributions may be noisy and produce unreliable or inconsistent results.

When comparing the distribution of Diagnostic + to Diagnostic - SNPs indicative of the At subgenome overwriting the Dt subgenome through hGC or HE (Figure 5C), we found that the total distribution of Diagnostic + SNPs was statistically dissimilar from that of Diagnostic - SNPs for all species ($P < 2.14 \times 10^{-17}$ for all species, two-sided Kolmogorov-Smirnov test); however, because we found that an excess of Diagnostic - SNPs in every species and for every category of SNP, this difference was not due to the presence of hGC or HE. For potential hGC or HE events in the opposite direction (i.e., in which the Dt subgenome overwrites the At subgenome), we found similar, though not as striking, patterns with the exclusion of AD4. These distributions have varying statistical significance in each species ($P = 1.8 \times 10^{-9}$ in *G. mustelinum*; $P = 1.8 \times 10^{-7}$ in *G. tomentosum*; $P = 3.4 \times 10^{-3}$ in *G. stephensii*;

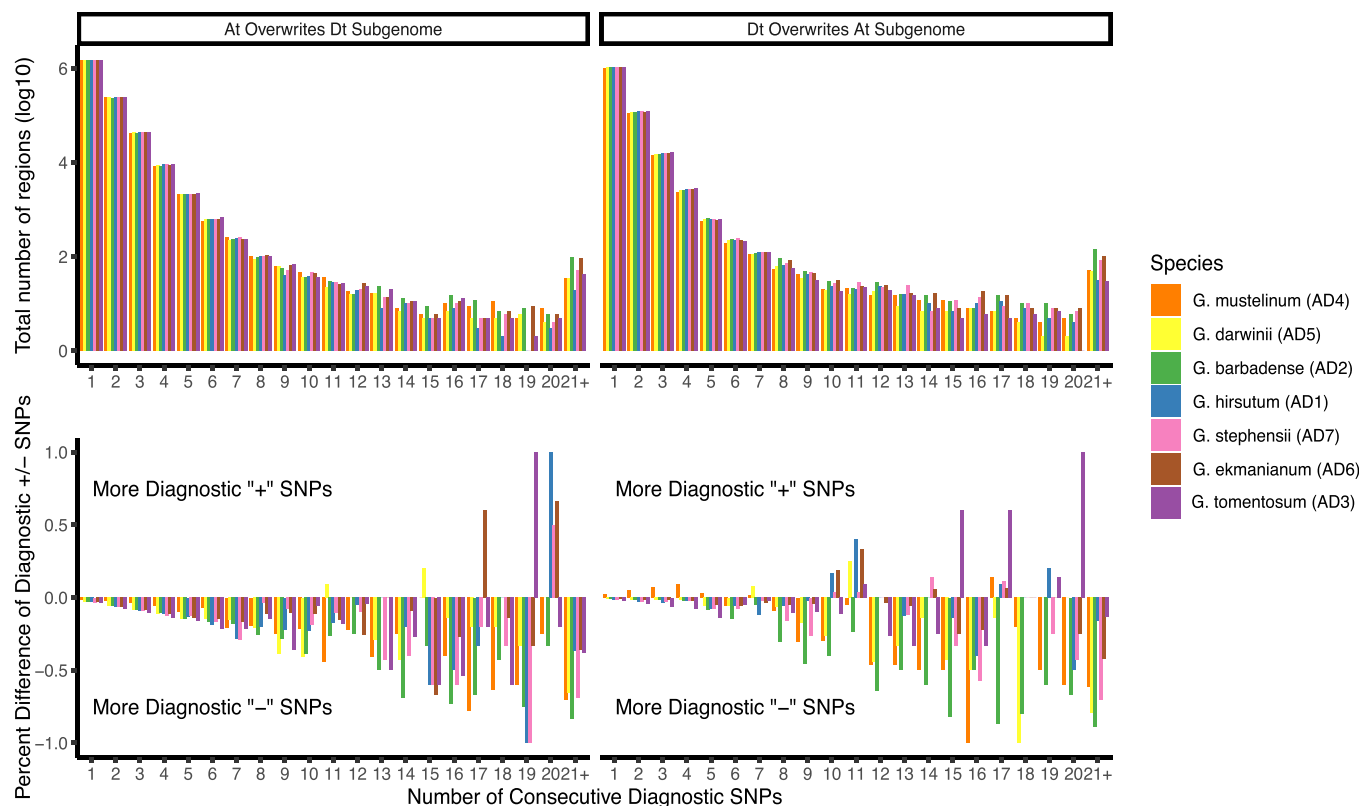


FIGURE 5 Classical “quartet” method fails to detect homoeologous gene conversion and homoeologous exchanges. Regions of the genome indicative of the Dt subgenome overwriting the At subgenome (A, C) or the reciprocal direction (B, D) identified using a modification of the classic “quartet” method. Regions were identified by first identifying homoeoSNPs (i.e., SNPs in which a subgenome, and its most closely related diploid progenitor have one allele, and the other subgenome and its most closely related diploid progenitor have a different allele at that site). The number of Diagnostic + or Diagnostic – SNPs located between two flanking homoeoSNPs is tabulated along the *x*-axis of each plot. (A, B) Total number of regions (*y*-axis) with a given number of consecutive Diagnostic + and – SNPs (*x*-axis). (C, D) The percentage difference between the number of regions with Diagnostic + or – SNPs, calculated as (difference between Diagnostic + and Diagnostic – regions/total number of regions). Positive values indicate more regions with Diagnostic + SNPs, negative values indicate more regions with Diagnostic – SNPs. Each polyploid is represented by a different color in the bar graph, shown at right. Although the classical “quartet” method only required two consecutive Diagnostic + SNPs to be inferred as hGC, we compared the total number of regions with the same number of Diagnostic + or – SNPs to infer hGC events.

$P = 3.9e-3$ in *G. hirsutum*; $P = 0.256$ in *G. barbadense*; $P = 0.413$ in *G. ekmanianum*; $P = 0.664$ in *G. darwinii*; two-sided Kolmogorov-Smirnov). In *G. mustelinum* (AD4), we found a slight excess of Diagnostic + SNP regions of size 1–5, but all other species had a deficit of Diagnostic + SNP regions of this size. For some of the larger values of consecutive SNPs, we saw that some species had an excess of Diagnostic + SNP regions, but there was no clear trend for hGC or HE in either direction, including the previously described HE event on chromosome 1. Thus, we conclude that quartet-based methods are unreliable for identifying either small regions of homoeologous gene conversion or larger homoeologous exchanges in allopolyploid cottons.

DISCUSSION

Genomes of polyploid species are notoriously difficult to analyze, due in no small part to the physical interactions between duplicated chromosomes in meiosis that lead to homoeologous exchanges (HEs) or homoeologous gene

conversion (hGC). Here, we developed a more robust and analytical framework for identifying the regions of an allopolyploid genome that have experienced these interactions, using a monophyletic group of seven allopolyploid cotton species with some previously described instances of hGC, but no documented cases of HEs, combined with five genomes of closely related diploid species. While previous analyses using “quartet”-based approaches (i.e., those using only the two subgenomes of an allotetraploid and a closely related diploid to each subgenome) estimated that between 1% and 7% of genes in allopolyploid *G. hirsutum* and *G. barbadense* have experienced hGC (Salmon et al., 2010; Flagel et al., 2012), our analytical method reveals that this range is a vast overestimation, suggesting that two and zero genes have potentially been influenced by hGC in *G. hirsutum* and *G. barbadense*, respectively. In the other five allotetraploid species in the clade, nearly 100 genes may have experienced hGC in total, but because nearly all of these events were not shared between any species (even between extremely closely related species, as is the case with *G. ekmanianum* and *G. stephensii*) and are located in

regions with poor assembly quality, it is difficult to rule out assembly or alignment artifacts as the cause of these patterns. In total, our results suggest that traditional “quartet”-based methods of inferring hGCs may be unreliable and produce dramatic overestimates of the rates of HE/hGC.

We also present the first evidence of an HE event within cotton genomes, affecting the terminal 175 kb (and 23 genes) of chromosome 1 in six of the seven genomes analyzed (although the seventh species has a poorer assembly at this region in chromosome 1 of both subgenomes, indicating that it has also likely experienced this HE event). This HE event is notable because the two-fold difference in size between homoeologous chromosomes has generally been assumed to restrict the pairing behavior of homoeologs during meiosis. As such, no multivalent chromosome pairing behavior has been observed in any of the allotetraploid cotton species. In fact, it has long been established (Beasley, 1940; Endrizzi, 1962) that intergenomic A genome \times D genome diploid hybrids in *Gossypium* exhibit only about six bivalents (of the 13 possible), whereas both natural and synthetic allopolyploids have nearly complete homoeologous pairing (Endrizzi, 1962). Moreover, chromosome pairing in synthetic allopolyploids is limited almost entirely to homoeologs. Thus, as noted by Endrizzi et al. (1985) and others, the restriction of pairing to homologous chromosomes that we see today most likely existed even at initial allopolyploid formation.

The fact that we see very low rates of HE correlated with low rates of hGC is probably not coincidental, as the same meiotic machinery and pathway that generates HEs is also responsible for generating hGC. Thus, while the end result of hGCs and HEs may be dissimilar in the direct meiotic products, their long-term genomic signatures may be difficult to distinguish from each other, particularly after several generations in which regions of HEs can be broken up via homologous recombination. Thus, further development of methods is warranted to identify hGC in species with documented examples of HEs and especially to distinguish hGC from HE. Although the SNP patterns in regions that have experienced HE should be identical to those that have experienced hGC, other attributes of these sites not explored here (for example, analyzing the distribution of hGC or HE tract length sizes compared to the distribution of haplotype block sizes genome-wide, or through the direct tracking of recombination events through the use of ancestral recombination graphs; Hayman et al., 2023) offer a potential opportunity to distinguish these intertwined phenomena. Analytical methods to evaluate subgenome architecture (Session and Rokhsar, 2023) and population-level processes in allotetraploids (e.g., demographic history, interploidy introgression) has received recent attention (Blischak et al., 2023; Wang et al., 2023), and we suggest that studying hGC and HE at the population level may offer additional insights that are not possible with our phylogenetic SNP-based approach.

While many studies have suggested that hGC could be a pathway for allopolyploid genomes to overcome genetic

incompatibilities or create new haplotypes on which selection can act (Gong et al., 2012, 2014; Sharbrough et al., 2017), our results suggest that because these studies are based on the “quartet” approach, their interpretation of identifying bona fide hGC regions should be treated with caution. While we do not disagree that, conceptually, hGCs can create novel haplotypes that might be subject to selection, it is important to robustly and systematically test this hypothesis before concluding that a particular region of an allopolyploid genome has experienced hGC and that selection has acted upon these regions. Ultimately, the detection of hGC and inferences of selection are attempts to attribute evolutionary processes to explain a pattern observed in data; as we show here, however, multiple evolutionary processes may generate the same patterns of data.

Although our current method is unable to identify with certainty which regions of the genome have experienced hGC (only the proportion of each SNP pattern described above that are overrepresented compared to their non-hGC counterparts), it is important to note that other meiotic recombination patterns, such as double crossovers between homoeologous chromosomes, may produce SNP patterns that appear similar to longer hGC tracts, but would only affect the region between the two breakpoints, rather than the regions that would be affected by HE (which affects the entirety of the telomere to the recombination breakpoint). However, because the typical tract length of hGCs or double crossovers involving homoeologous chromosomes is not known, it is not possible to distinguish these two processes, which remains a topic for future work, presumably in a system that has more easily identifiable regions that have experienced hGC, HE, and double crossovers. Analytical methods to identify the frequency of occurrence and the allelic segregation patterns of double crossovers is an active area of research in autopolyploids (Griswold and Williamson, 2017; Griswold and Asif, 2023), but extending these methods to allopolyploids that experience HEs or hGC has received little attention.

Extending this analytical pipeline to other allopolyploid systems will be of interest to others, so it is germane to consider some of the necessary criteria for taxon selection. The most important aspect of choosing species for this analysis is to identify those with minimal amounts of gene flow, so that the SNP patterns are not influenced by this inherently homogenizing effect. The species in *Gossypium* used here are probably an outlier in this respect, as most diploid progenitors to an allopolyploid are probably not distributed in different hemispheres separated by an ocean. Additionally, the availability and relative divergence of diploid outgroups matters because it influences the total number and relative symmetry of SNP sites. We saw a marked increase between the total number of SNP sites in the A lineage that were attributable to either ILS, recurrent mutations, or back mutations, compared to the D lineage. This increase is likely due to the phylogenetic position of the diploid outgroups, where the diploid outgroup for the A lineage (F1) was considerably more distant to the polyploid subgenome than the D lineage

outgroup (D10). Hence, there are more opportunities for recurrent mutations and back mutations to occur on the terminal branches of F1 and either of the terminal branches of A2 or the At subgenome. Although we did not directly test the cause of this observation, it suggests that ILS was responsible for a comparatively small portion of the SNPs used here and that the majority of the Diagnostic – SNPs may be caused by recurrent or back mutations. Thus, when choosing a diploid outgroup, there is a trade-off between choosing a species that is distantly related enough to minimize the amount of ILS, with one that is close enough to minimize the amount of recurrent or back mutation. Finally, we note the requirement for high-quality genome assemblies for an allopolyploid (and the increase in interpretive possibility from including more than one) and for genomes for both model diploid genome donors and their outgroups and to the system overall. Although these are relatively stringent methodological requirements, the number of high-quality genomes continues to increase in recent years, and genomic tools such as whole-genome alignment algorithms designed specifically for the complexities of plant genomes are under active development (Song et al., 2023). Thus, the application of methods such as those presented here may provide additional insight into the genomic location, rate, and tempo of hGC and HE events and their consequences for selection and adaptation in allopolyploids.

AUTHOR CONTRIBUTIONS

J.L.C. developed the methodology, analyzed data, acquired funding, and wrote the original draft. J.S. analyzed data and acquired funding. C.E.G., D.B.S., D.G.P., and J.F.W. acquired funding. J.L.C., C.E.G., J.S., D.B.S., and J.F.W. conceptualized the project. All authors reviewed and edited the manuscript.



ACKNOWLEDGMENTS

We thank ResearchIT at Iowa State University for computational support and two anonymous reviewers for helpful comments that improved the manuscript. This work was supported by a National Science Foundation Post-doctoral Research Fellowship in Biology (IOS-2209085 to J.L.C.), and National Science Foundation (NSF) awards IOS-1829176 (D.B.S., J.F.W., J.S., and C.E.G.), IOS-2145811 (J.S.), and U.S. Department of Agriculture ARS 58-6066-0-066 (D.G.P., C.E.G., and J.F.W.). Open access funding provided by the Iowa State University Library.

DATA AVAILABILITY STATEMENT

Scripts for all genome alignments and data filtration are available on Github (<https://github.com/conJUSTover/GeneConversion>), and raw alignment and filtered VCF files are available on Figshare (<https://doi.org/10.25422/azu.data.24512896>).

ORCID

Justin L. Conover  <http://orcid.org/0000-0002-3558-6000>
Corrinne E. Grover  <http://orcid.org/0000-0003-3878-5459>

Joel Sharbrough  <http://orcid.org/0000-0002-3642-1662>
Daniel B. Sloan  <http://orcid.org/0000-0002-3618-0897>
Daniel G. Peterson  <http://orcid.org/0000-0002-0274-5968>
Jonathan F. Wendel  <http://orcid.org/0000-0003-2258-5081>

REFERENCES

- Akagi, T., K. Jung, K. Masuda, and K. K. Shimizu. 2022. Polyploidy before and after domestication of crop species. *Current Opinion in Plant Biology* 69: 102255.
- Beasley, J. O. 1940. The origin of American tetraploid *Gossypium* species. *American Naturalist* 74: 285–286.
- Bird, K. A., C. E. Niederhuth, S. Ou, M. Gehan, J. C. Pires, Z. Xiong, R. VanBuren, and P. P. Edger. 2021. Replaying the evolutionary tape to investigate subgenome dominance in allopolyploid *Brassica napus*. *New Phytologist* 230: 354–371.
- Bird, K. A., J. C. Pires, R. VanBuren, Z. Xiong, and P. P. Edger. 2023. Dosage-sensitivity shapes how genes transcriptionally respond to allopolyploidy and homoeologous exchange in resynthesized *Brassica napus*. *Genetics* 225: iyad114.
- Blischak, P. D., M. Sajan, M. S. Barker, and R. N. Gutenkunst. 2023. Demographic history inference and the polyploid continuum. *Genetics* 224: iyad107.
- Bradbury, P. J., T. Casstevens, S. E. Jensen, L. C. Johnson, Z. R. Miller, B. Monier, M. C. Romy, et al. 2022. The Practical Haplotype Graph, a platform for storing and using pangenomes for imputation. *Bioinformatics* 38: 3698–3702.
- Chaudhary, B., R. Hovav, R. Rapp, N. Verma, J. A. Udall, and J. F. Wendel. 2008. Global analysis of gene expression in cotton fibers from wild and domesticated *Gossypium barbadense*. *Evolution & Development* 10: 567–582.
- Chen, Z. J., A. Sreedasyam, A. Ando, Q. Song, L. M. De Santiago, A. M. Hulse-Kemp, M. Ding, et al. 2020. Genomic diversifications of five *Gossypium* allopolyploid species and their impact on cotton improvement. *Nature Genetics* 52: 525–533.
- Chester, M., J. P. Gallagher, V. V. Symonds, A. V. Cruz da Silva, E. Mavrodiev, A. R. Leitch, P. S. Soltis, and D. E. Soltis. 2012. Extensive chromosomal variation in a recently formed natural allopolyploid species, *Tragopogon miscellus* (Asteraceae). *Proceedings of the National Academy of Sciences, USA* 109: 1176–1181.
- Conover, J. L., and J. F. Wendel. 2022. Deleterious mutations accumulate faster in allopolyploid than diploid cotton (*Gossypium*) and unequally between subgenomes. *Molecular Biology and Evolution* 39: msac024.
- Cronn, R., R. L. Small, T. Haselkorn, and J. F. Wendel. 2003. Cryptic repeated genomic recombination during speciation in *Gossypium gossypoides*. *Evolution* 57: 2475–2489.
- Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, et al. 2011. The variant call format and VCFtools. *Bioinformatics* 27: 2156–2158.
- Deb, S. K., P. P. Edger, J. C. Pires, and M. R. McKain. 2023. Patterns, mechanisms, and consequences of homoeologous exchange in allopolyploid angiosperms: a genomic and epigenomic perspective. *New Phytologist* 238: 2284–2304.
- Doyle, J. J., and A. N. Egan. 2010. Dating the origins of polyploidy events. *New Phytologist* 186: 73–85.
- Edger, P. P., M. R. McKain, K. A. Bird, and R. VanBuren. 2018. Subgenome assignment in allopolyploids: challenges and future directions. *Current Opinion in Plant Biology* 42: 76–80.
- Endrizzi, J. E. 1962. The diploid-like cytological behavior of tetraploid cotton. *Evolution* 16: 325–329.
- Endrizzi, J. E., E. L. Turcotte, and R. J. Kohel. 1985. Genetics, cytology, and evolution of *Gossypium*. In E. W. Caspari and J. G. Scandalios [eds.], *Advances in Genetics*, vol. 23, 271–375. Academic Press, Cambridge, MA, USA.

- Flagel, L. E., J. F. Wendel, and J. A. Udall. 2012. Duplicate gene evolution, homoeologous recombination, and transcriptome characterization in allopolyploid cotton. *BMC Genomics* 13: 302.
- Fryxell, P. A. 1992. A revised taxonomic interpretation of *Gossypium* L. (Malvaceae). *Rheedea: Official Journal of Indian Association for Angiosperm Taxonomy* 2: 108–116.
- Gaeta, R. T., and J. C. Pires. 2010. Homoeologous recombination in allopolyploids: the polyploid ratchet. *New Phytologist* 186: 18–28.
- Gaeta, R. T., J. C. Pires, F. Iniguez-Luy, E. Leon, and T. C. Osborn. 2007. Genomic changes in resynthesized *Brassica napus* and their effect on gene expression and phenotype. *Plant Cell* 19: 3403–3417.
- Gong, L., M. Olson, and J. F. Wendel. 2014. Cytonuclear evolution of rubisco in four allopolyploid lineages. *Molecular Biology and Evolution* 31: 2624–2636.
- Gong, L., A. Salmon, M.-J. Yoo, K. K. Grupp, Z. Wang, A. H. Paterson, and J. F. Wendel. 2012. The cytonuclear dimension of allopolyploid evolution: an example from cotton using rubisco. *Molecular Biology and Evolution* 29: 3023–3036.
- Griswold, C. K., and S. Asif. 2023. Meiosis at three loci in autotetraploids: Probabilities of gamete modes and genotypes without and with preferential cross-over formation. *Heredity* 130: 223–235.
- Griswold, C. K., and M. W. Williamson. 2017. A two-locus model of selection in autotetraploids: chromosomal gametic disequilibrium and selection for an adaptive epistatic gene combination. *Heredity* 119: 314–327.
- Grover, C. E., M. A. Arick 2nd, A. Thrash, J. L. Conover, W. S. Sanders, D. G. Peterson, J. E. Frelichowski, et al. 2019. Insights into the evolution of the New World diploid cottons (*Gossypium*, subgenus *Houzingenia*) based on genome sequencing. *Genome Biology and Evolution* 11: 53–71.
- Grover, C. E., K. K. Grupp, R. J. Wanzek, and J. F. Wendel. 2012. Assessing the monophyly of polyploid *Gossypium* species. *Plant Systematics and Evolution = Entwicklungsgeschichte und Systematik der Pflanzen* 298: 1177–1183.
- Grover, C. E., M. Pan, D. Yuan, M. A. Arick, G. Hu, L. Brase, D. M. Stelly, et al. 2020. The *Gossypium longicalyx* genome as a resource for cotton breeding and evolution. *G3 Genes[Genomes]Genetics* 10: 1457–1467.
- Guo, H., X. Wang, H. Gundlach, K. F. X. Mayer, D. G. Peterson, B. E. Scheffler, P. W. Chee, and A. H. Paterson. 2014. Extensive and biased intergenomic nonreciprocal DNA exchanges shaped a nascent polyploid genome, *Gossypium* (cotton). *Genetics* 197: 1153–1163.
- Hayman, E., A. Ignatieva, and J. Hein. 2023. Recoverability of ancestral recombination graph topologies. *Theoretical Population Biology* 154: 27–39.
- Hu, G., C. E. Grover, D. Yuan, Y. Dong, E. Miller, J. L. Conover, and J. F. Wendel. 2021. Evolution and diversity of the cotton genome. In M.-U.- Rahman, Y. Zafar, and T. Zhang [eds], *Cotton precision breeding*, 25–78. Springer International, Cham, Switzerland.
- Jiao, Y., N. J. Wickett, S. Ayyampalayam, A. S. Chanderbali, L. Landherr, P. E. Ralph, L. P. Tomsho, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473: 97–100.
- Kryvokhyzh, D., A. Salcedo, M. C. Eriksson, T. Duan, N. Tawari, J. Chen, M. Guerrina, et al. 2019. Parental legacy, demography, and admixture influenced the evolution of the two subgenomes of the tetraploid *Capsella bursa-pastoris* (Brassicaceae). *PLoS Genetics* 15: e1007949.
- Li, F., G. Fan, C. Lu, G. Xiao, C. Zou, R. J. Kohel, Z. Ma, et al. 2015. Genome sequence of cultivated upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. *Nature Biotechnology* 33: 524–530.
- Lim, K. Y., D. E. Soltis, P. S. Soltis, J. Tate, R. Matyasek, H. Srubarova, A. Kovarik, et al. 2008. Rapid chromosome evolution in recently formed polyploids in *Tragopogon* (Asteraceae). *PLoS One* 3: e3353.
- Li, N., C. Xu, A. Zhang, R. Lv, X. Meng, X. Lin, L. Gong, et al. 2019. DNA methylation repatterning accompanying hybridization, whole genome doubling and homoeolog exchange in nascent segmental rice allotetraploids. *New Phytologist* 223: 979–992.
- Liu, H., J. Huang, X. Sun, J. Li, Y. Hu, L. Yu, G. Liti, et al. 2018. Tetrad analysis in plants and fungi finds large differences in gene conversion rates but no GC bias. *Nature Ecology & Evolution* 2: 164–173.
- Lorenz, A., and S. J. Mpaulo. 2022. Gene conversion: a non-Mendelian process integral to meiotic recombination. *Heredity* 129: 56–63.
- Malinsky, M., M. Matschiner, and H. Svardal. 2021. Dsuite - Fast D-statistics and related admixture evidence from VCF files. *Molecular Ecology Resources* 21: 584–595.
- Mason, A. S., and J. F. Wendel. 2020. Homoeologous exchanges, segmental allopolyploidy, and polyploid genome evolution. *Frontiers in Genetics* 11: 1014.
- Meng, Q., J. Gu, Z. Xu, J. Zhang, J. Tang, A. Wang, P. Wang, et al. 2023. Comparative analysis of genome sequences of the two cultivated tetraploid cottons, *Gossypium hirsutum* (L.) and *G. barbadense* (L.). *Industrial Crops and Products* 196: 116471.
- Narasimhan, V., P. Danecsek, A. Scally, Y. Xue, C. Tyler-Smith, and R. Durbin. 2016. BCFtools/ROH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* 32: 1749–1751.
- One Thousand Plant Transcriptomes Initiative. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574: 679–685.
- Ortiz, A. J., and J. Sharbrough. 2023. Genome-wide patterns of homoeologous gene flow in allotetraploid coffee. *bioRxiv* 2023.09.10.557041 [preprint].
- Page, J. T., M. D. Huynh, Z. S. Liechty, K. Grupp, D. Stelly, A. M. Hulse, H. Ashrafi, et al. 2013. Insights into the evolution of cotton diploids and polyploids from whole-genome re-sequencing. *G3 Genes[Genomes]Genetics* 3: 1809–1818.
- Page, J. T., Z. S. Liechty, R. H. Alexander, K. Clemons, A. M. Hulse-Kemp, H. Ashrafi, A. Van Deynze, et al. 2016. DNA sequence evolution and rare homoeologous conversion in tetraploid cotton. *PLoS Genetics* 12: e1006012.
- Peng, R., Y. Xu, S. Tian, T. Unver, Z. Liu, Z. Zhou, X. Cai, et al. 2022. Evolutionary divergence of duplicated genomes in newly described allotetraploid cottons. *Proceedings of the National Academy of Sciences, USA* 119: e2208496119.
- Perkin, L. C., A. Bell, L. L. Hinze, C. P.-C. Suh, M. A. Arick, D. G. Peterson, and J. A. Udall. 2021. Genome assembly of two nematode-resistant cotton lines (*Gossypium hirsutum* L.). *G3 Genes[Genomes]Genetics* 11: jkab276.
- Renny-Byfield, S., and J. F. Wendel. 2014. Doubling down on genomes: Polyploidy and crop plants. *American Journal of Botany* 101: 1711–1725.
- Salmon, A., L. Flagel, B. Ying, J. A. Udall, and J. F. Wendel. 2010. Homoeologous nonreciprocal recombination in polyploid cotton. *New Phytologist* 186: 123–134.
- Session, A. M., and D. S. Rokhsar. 2023. Transposon signatures of allopolyploid genome evolution. *Nature Communications* 14: 3180.
- Sharbrough, J., J. L. Conover, M. Fernandes Gyorfy, C. E. Grover, E. R. Miller, J. F. Wendel, and D. B. Sloan. 2022. Global patterns of subgenome evolution in organelle-targeted genes of six allotetraploid angiosperms. *Molecular Biology and Evolution* 39: msac074.
- Sharbrough, J., J. L. Conover, J. A. Tate, J. F. Wendel, and D. B. Sloan. 2017. Cytonuclear responses to genome doubling. *American Journal of Botany* 104: 1277–1280.
- Song, B., E. S. Buckler, and M. C. Stitzer. 2023. New whole-genome alignment tools are needed for tapping into plant diversity. *Trends in Plant Science* 29: 355–369.
- Song, B., S. Marco-Sola, M. Moreto, L. Johnson, E. S. Buckler, and M. C. Stitzer. 2022. AnchorWave: Sensitive alignment of genomes with high sequence diversity, extensive structural polymorphism, and whole-genome duplication. *Proceedings of the National Academy of Sciences, USA* 119: e2113075119.
- Taghian, D. G., and J. A. Nickoloff. 1997. Chromosomal double-strand breaks induce gene conversion at high frequency in mammalian cells. *Molecular and Cellular Biology* 17: 6386–6393.
- Udall, J. A., E. Long, C. Hanson, D. Yuan, T. Ramaraj, J. L. Conover, L. Gong, et al. 2019a. De novo genome sequence assemblies of *Gossypium raimondii* and *Gossypium turneri*. *G3 Genes[Genomes]Genetics* 9: 3079–3085.

- Udall, J. A., E. Long, T. Ramaraj, J. L. Conover, D. Yuan, C. E. Grover, L. Gong, et al. 2019b. The genome sequence of *Gossypioides kirkii* illustrates a descending dysploidy in plants. *Frontiers in Plant Science* 10: 1541.
- Viot, C. R., and J. F. Wendel. 2023. Evolution of the cotton genus, *Gossypium*, and its domestication in the Americas. *Critical Reviews in Plant Sciences* 42: 1–33.
- Wang, K., J. F. Wendel, and J. Hua. 2018. Designations for individual genomes and chromosomes in *Gossypium*. *Journal of Cotton Research* 1: 3.
- Wang, M., J. Li, P. Wang, F. Liu, Z. Liu, G. Zhao, Z. Xu, et al. 2021. Comparative genome analyses highlight transposon-mediated genome expansion and the evolutionary architecture of 3D genomic folding in cotton. *Molecular Biology and Evolution* 38: 3621–3636.
- Wang, T., A. D. J. van Dijk, J. Bucher, J. Liang, J. Wu, G. Bonnema, and X. Wang. 2023. Interploidy introgression shaped adaptation during the origin and domestication history of *Brassica napus*. *Molecular Biology and Evolution* 40: msad199.
- Wendel, J., and J. Doyle. 2005. Polyploidy and evolution in plants. In R. J. Henry [ed.], *Plant diversity and evolution: genotypic and phenotypic variation in higher plants*, 97–117. CABI Publishing, Wallingford, UK.
- Wendel, J. F., and C. E. Grover. 2015. Taxonomy and evolution of the cotton genus, *Gossypium*. In D. D. Fang and R. G. Percy [eds.], *Cotton*, 2nd ed., vol. 57, 25–44. American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America, Madison, WI, USA.
- Wendel, J. F., A. Schnabel, and T. Seelanan. 1995. An unusual ribosomal DNA sequence from *Gossypium gossypoides* reveals ancient, cryptic, intergenomic introgression. *Molecular Phylogenetics and Evolution* 4: 298–313.
- Wu, T. D., J. Reeder, M. Lawrence, G. Becker, and M. J. Brauer. 2016. GMAP and GSNAP for genomic sequence alignment: enhancements to speed, accuracy, and functionality. *Methods in Molecular Biology* 1418: 283–334.
- Wu, Y., F. Lin, Y. Zhou, J. Wang, S. Sun, B. Wang, Z. Zhang, et al. 2021. Genomic mosaicism due to homoeologous exchange generates extensive phenotypic diversity in nascent allopolyploids. *National Science Review* 8: nwaa277.
- Yu, J., S. Jung, C.-H. Cheng, T. Lee, P. Zheng, K. Buble, J. Crabb, et al. 2021. CottonGen: The community database for cotton genomics, genetics, and breeding research. *Plants* 10: 2805.
- Yuan, D., C. E. Grover, G. Hu, M. Pan, E. R. Miller, J. L. Conover, S. P. Hunt, et al. 2021. Parallel and intertwining threads of domestication in allopolyploid cotton. *Advancement of Science* 8: 2003634.
- Zhao, J., J. Li, R. Lv, B. Wang, Z. Zhang, T. Yu, S. Liu, et al. 2023. Meiotic pairing irregularity and homoeologous chromosome compensation cause rapid karyotype variation in synthetic allotetraploid wheat. *New Phytologist* 239: 606–623.
- Zhao, X. P., Y. Si, R. E. Hanson, C. F. Crane, H. J. Price, D. M. Stelly, J. F. Wendel, and A. H. Paterson. 1998. Dispersed repetitive DNA has spread to new genomes since polyploid formation in cotton. *Genome Research* 8: 479–492.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Appendix S1. Supplementary figures.

Appendix S2. Supplementary analysis of introgression in *Gossypium*.

Appendix S3. Supplementary tables.

How to cite this article: Conover, J. L., C. E. Grover, J. Sharbrough, D. B. Sloan, D. G. Peterson, and J. F. Wendel. 2024. Little evidence for homoeologous gene conversion and homoeologous exchange events in *Gossypium* allopolyploids. *American Journal of Botany* 111(8): e16386. <https://doi.org/10.1002/ajb2.16386>