the geneticssociety

ARTICLE

Comparative transmission genetics of introgressed chromatin in reciprocal advanced backcross populations in *Gossypium* (cotton) polyploids

Jeevan Adhikari[1], Rahul Chandnani[2], Deepak Vitrakoti[1], Sameer Khanal[3], Wiriyanat Ployaram[1] and Andrew H. Paterson[1]✉

Introgression is a potential source of valuable genetic variation and interspecific introgression lines are important resources for plant breeders to access novel alleles. Experimental advanced-generation backcross populations contain individuals with genomic compositions similar to those resulting from natural interspecific hybridization and provide opportunities to study the nature and transmission pattern of donor chromatin in recipient genomes. Here, we analyze transmission of donor chromatin in reciprocal backcrosses between *G. hirsutum* and *G. barbadense*. Across the genome, recurrent backcrossing in both backgrounds yielded donor chromatin at slightly higher frequencies than the Mendelian expectation in BC_5F_1 plants, while the average frequency of donor alleles in BC_5F_2 segregating families was less than expected. In the two subgenomes of polyploid cotton, the rate of donor chromatin introgression was similar. Although donor chromatin was tolerated over much of the recipient genomes, 21 regions recalcitrant to donor alleles were identified. Only limited correspondence is observed between the recalcitrant regions in the two backgrounds, suggesting the effect of species background on introgression of donor segments. Genetic breakdown was progressive, with floral abscission and seed inviability ongoing during backcrossing cycles. Regions of either high or low introgression tended to be in terminal chromosomal regions that are generally rich in both genes and crossover events, with long stretches around the centromere having limited crossover activity resulting in relatively constant low introgression frequencies. Constraints on fixation and selection of donor alleles highlights the challenges of utilizing introgression breeding in crop improvement.

Heredity (2023) 130:209–222; https://doi.org/10.1038/s41437-023-00594-w

INTRODUCTION

Interspecific gene transfer is a potential source of valuable genetic variation, and interspecific hybridization has been an attractive natural means for introducing novel and selectable variation for important traits into crop improvement (Anderson 1949; Levi et al. 2009; Tanksley and Nelson 1996). Gene flow via interspecific hybridization can provide raw material for natural selection and evolutionary change. Introgression of chromosomal segments is one of the consequences of interspecific hybridization, which may result from backcrossing following the initial hybridization (Grant 1981). In addition to introducing genes for adaptive traits (Heiser 1979; Waghmare et al 2016), introgression can reduce reproductive isolation barriers (Meyn and Emboden 1987) and broaden the genetic base of a crop species by incorporating novel alleles / allele combinations (Adhikari et al. 2017; Paterson et al. 2004). While gene flow via hybridization and introgression can be a significant substrate for evolution (Anderson 1949), genomic regions acting as barriers to gene flow are important for species integrity. As such, identification and investigation of such regions might shed light on factors responsible for reproductive isolation (Baack et al. 2015).

Experimental advanced-generation backcross populations contain individuals with genomic compositions resembling those resulting from natural interspecific hybridization. Individual members of such advanced-generation populations usually retain some genomic features of the donor parent while they most closely resemble their recurrent (backcross) parent. Experimental introgression populations are important genetic resources not only for crop improvement but also to study gene flow between species (Jiang et al. 2000). Although introgression has been widely acknowledged as a potential source of valuable genetic variation to enrich crop gene pools, it has had varying and often limited effect in practice (Hajjar and Hodgkin 2007) due to limited availability of genetic markers and genetic resources in the past. Availability of genomic resources such as reference genomes and abundant generation of genetic markers at modest cost have facilitated the study of gene flow among populations (Kim et al. 2016; Paterson et al. 2012).

Cotton belongs to the genus *Gossypium*, comprised of more than 50 species, of which about 45 are diploid ($2n = 26$) and 7 are allotetraploid ($2n = 4\times = 52$). In addition to being an important economic crop and leading textile fiber, cotton is well suited for studies of introgressive hybridization and in particular the influences of polyploidy on levels and patterns of introgression. Tetraploids contain two distinct subgenomes—the At subgenome resembles the extant A genome of *G. herbaceum* L and the Dt subgenome

[1]Plant Genome Mapping Laboratory, Department of Crop and Soil Sciences, University of Georgia, Athens, GA, USA. [2]Global Institute for Food Security, Saskatoon, SK, Canada. [3]Institute of Plant Breeding Genetics and Genomics, Department of Crop and Soil Sciences, University of Georgia, Tifton, GA, USA. Associate editor: Lindsey Compton
✉email: paterson@uga.edu

Received: 12 August 2022 Revised: 11 January 2023 Accepted: 23 January 2023
Published online: 8 February 2023

resembles the D genome of *G. raimondii* Ulbrich or *G. gossypoides* Ulbrich (Wendel et al. 1995). The A- and D-genome species are estimated to have diverged from a common ancestor 6–11 million years ago (mya) and hybridized (followed by polyploidization) about 1–2 mya (Wendel 1989). After polyploidization, several chromosomal rearrangements occurred distinguishing the tetraploid (AD) genomes from their diploid progenitors (Brubaker et al. 1999; Desai et al. 2006; Rong et al. 2004). Although normal meiotic chromosome pairing has suggested little structural rearrangement since the divergence of *G. hirsutum* and *G. barbadense* (Beasley 1942), comprehensive linkage and genetic maps (Rong et al. 2004; Waghmare et al. 2005; Yu et al. 2007) have suggested some possible small rearrangements among the chromosomes of these two tetraploid Gossypium species. These comprehensive linkage maps and high-contiguity genome sequences (Paterson et al. 2012; Zhang et al. 2015) provided for detailed study of *Gossypium* transmission genetics.

Introgression and retention patterns of *G. barbadense*, *G. tomentosum* and *G. mustelinum* chromosome segments in *G. hirsutum* background have been studied previously (Jiang et al. 2000; Waghmare et al. 2016), with multilocus interactions suggested to play major roles in determination of the genomic composition of populations. Large and widespread deficiencies of donor (*G. barbadense*) chromatin were found, including seven independent chromosomal regions showing no introgression into *G. hirsutum*. The At and Dt subgenomes of allotetraploid cotton have been suggested to play different roles in evolution and consequently differ in retention of donor chromatin.

Both *G. hirsutum* and *G. barbadense* are cultivated tetraploid species that originated from a common ancestor about 1–2 million years ago and are cross compatible. However, during their evolution, they have accumulated reproductive barriers indicated by hybrid breakdown in interspecific F_2 and advanced generations (Zhang et al. 2014). Hybrid breakdown, instability, and selective elimination of desirable genes during selfing have been reported as the major obstacles for successful introgression breeding. While hybrid breakdown is poorly understood, some probable causes have been hypothesized including zygotic selection by duplicate recessive complementary genes for traits such as chlorophyll deficiency, asynapsis, corky and open buds. Chromosomal differences between Pima and Upland cotton may also cause gamete selection including translocations, minor inversion, and other cryptic structural differences. These events might result in pollen sterility or inhibition on pollen tube growth, suppression of recombination or crossover, selective elimination of genes and segregation distortions. Kantartzi and Roupakias (2008) showed that pollen tubes grew in a crooked manner in interspecific hybrids while they grew normally in intraspecific hybrids, and they found several forms of pollen tube inhibition in interspecific hybrids that were not seen in intraspecific hybrids. While interspecific hybridization is an important source of genetic variation, one of the possible strategies to minimize hybrid breakdown in interspecific hybrids is to stabilize the interspecific genetic background. While a doubled haploid strategy via semigamy might be used (Zhang and Stewart 2004), large-scale haploid production and chromosome doubling technique still remains to be established for cotton. Alternatives include the production of Pima cotton chromosome substitution lines (CSLs) or chromosome segment substitution lines (CSSLs) in an Upland cotton background or development of an advanced backcross population to select chromosomal segment introgression lines (CSILs) or near-isogenic lines (NILs) in a recurrent parent background.

Although *G. barbadense* introgression into *G. hirsutum* has been studied (Jiang et al. 2000), *G. hirsutum* introgression into *G. barbadense* background remains to be explored extensively. A survey of elite genotypes revealed five genomic regions of prominent historical introgression of *G. hirsutum* chromatin into *G. barbadense* (Wang et al. 1995), but provides no information about early generations or whether these introgressions were related to natural differences between the taxa or selection for a trait(s) by plant breeders. A detailed genetic recombination map of cotton provided further insights into the transmission genetics of *G. hirsutum* into *G. barbadense* (Rong et al. 2004) in addition to features of genome organization and evolution of cotton.

In this study we examine the transmission genetics of advanced-generation backcross progenies and resulting near isogenic lines developed from a cross between *Gossypium hirsutum* L. and *G. barbadense* L. In this paper we address the levels and patterns of introgression and retention of donor chromatin in the recurrent genome after several generations of backcrossing. We show that these cultivated species have differential introgression permeability and donor genome retention. Segregation patterns across genomes provide insights into reproductive barriers that affect both natural populations and crop gene pools. We also investigate segregation pattern of introgressed alleles and their deviation from expected Mendelian ratios. The segregation distorted regions (SDRs) identified based on these segregation patterns and the availability of reference genome enabled us to study gene family enrichment in genomic regions that are significantly resistant to introgression and might be important in species isolation. This study contributes to understanding gene flow between cultivated species of cotton and provides a platform for hypotheses about possible roles of specific genomic regions or genes that influence genome composition of these species.

MATERIALS AND METHODS

Plant materials and population development

Plant materials used in this study were developed from a set of reciprocal crosses between *Gossypium hirsutum* acc. Acala Maxxa and *G. barbadense* acc. Pima S6 (both inbred lines). These genotypes have been extensively used to produce molecular tools and resources including BAC libraries and Illumina genome sequences. Reciprocal advanced backcross populations were developed by first crossing the parents reciprocally (Acala Maxxa (♀) × Pima S6 (♂)—hereafter referred to as *G. hirsutum* background; and Pima S6 (♀) × Acala Maxxa (♂)—hereafter referred to as *G. barbadense* background), then independently backcrossing F_1 plants to the respective maternal parent to create 300–400 BC_1 progenies for each cross. The backcrossing scheme included planting only one seed from each preceding backcross to generate the next generation (Fig. 1). After five generation of backcrossing, 179 BC_5F_1 plants from the *G. hirsutum* background and 190 BC_5F_1 plants from *G. barbadense* were self-pollinated and a total of 8364 BC_5F_2 plants (2–32 individuals in each BC_5F_2 family) were grown at Iron Horse Farm, Watkinsville, Georgia in 2019 under cultural conditions consistent with commercial irrigated cotton production.

Genotyping

The genomic composition of the BC_5F_1 plants was inferred based on genotyping by sequencing (GBS). DNA was extracted from the parents and 369 BC_5F_1 plants using a scaled-down version of a published CTAB protocol (Paterson et al. 1993). A total of five multiplexed GBS libraries were constructed according to Andolfatto et al. (2011) wherein the DNA were double digested with HinP1I-HaeIII enzymes. The libraries were sequenced on Illumina MiSeq (in-house) with 75 bp single end reads (SE75). The TASSEL5 GBSV2 pipeline was used for sequence data processing and genotype calling (Glaubitz et al. 2014). Reads were aligned to *G. hirsutum* acc. *TM-1* (Zhang et al. 2015) using Burrow-Wheeler Alignment (bwa) and exported to variant call format (VCF). To minimize sequencing errors, only the first 64 base pairs were used to map reads to the reference genome. Filtering of the VCF was done for bi-allelic SNPs using Fisher's exact test with a threshold P value < 0.001, considering that true variants should represent biallelic homozygous state for inbred accessions. Genotypes for lines in *G. hirsutum* background were called together and those for lines in *G. barbadense* background cross were also called together. The SNPs were filtered for MAF > 0.01, missing <30% and heterozygous <10% at the population level. The retained SNPs were imputed using the Fast Inbred Line Library Imputation (FILLIN) pipeline available in TASSEL5 GBSv2 (Kelly et al. 2014).

The genomic composition of BC_5F_2 plants were inferred based on targeted microsatellite (SSR) genotyping of the introgressed chromosomal segments identified in their respective BC_5F_1 parents. At least two (and at most four) SSR markers were used to verify most of the introgressed

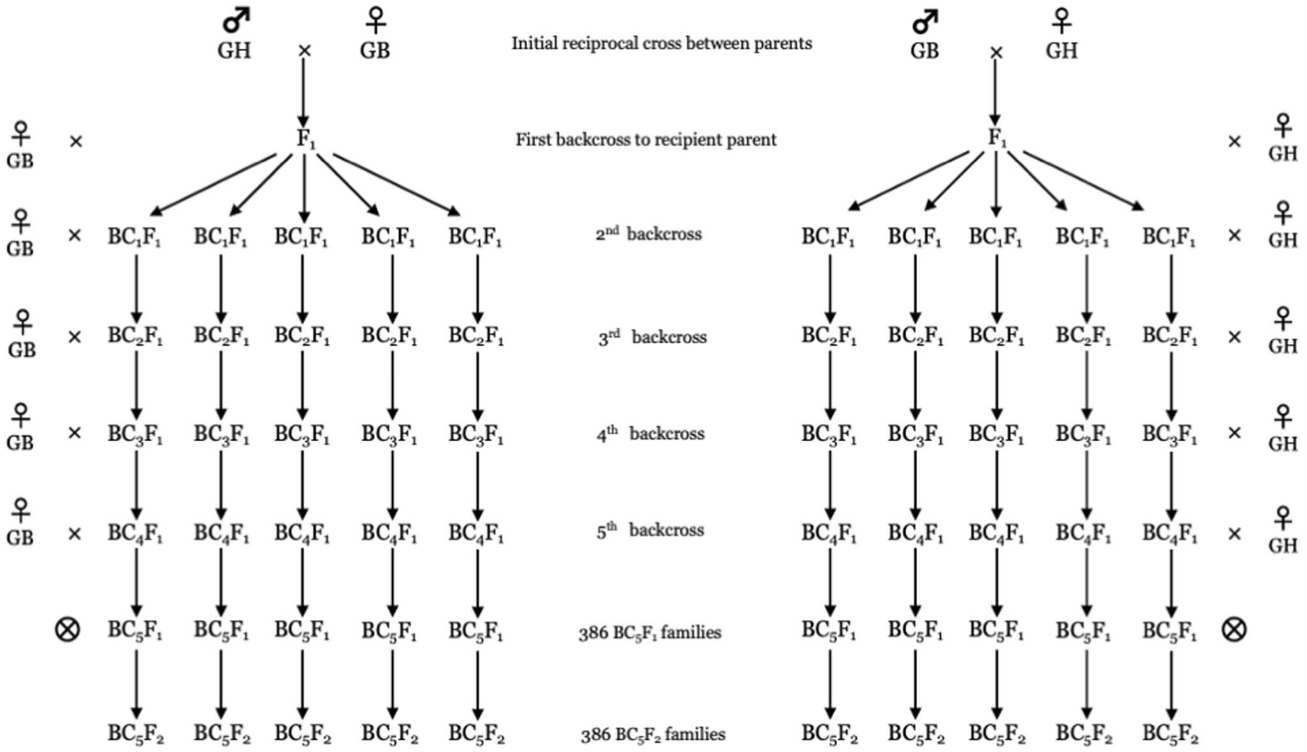

Fig. 1 Development of reciprocal set of advanced-backcross populations. GH and GB denote *G. hirsutum* and *G. barbadense* respectively. Each backcross lineage was advanced by single-seed descent.

regions while for small introgressions only one SSR marker was deployed. A total of 852 polymorphic SSR markers spanning the introgressed regions were derived from several published genetic maps of crosses between *G. barbadense* and *G. hirsutum* stored in the CottonGen SSR database (https://www.cottongen.org/data/download/marker). A total of 47 candidate SSRs were monomorphic in our lines and discarded, as were 23 with ambiguous bands. Among the 8364 BC_5F_2 individuals planted in 2019, the remaining 782 SSR markers were used to genotype 5315 plants (from BC_5F_1 parents carrying 2 to 5 introgressions) for the presence (or absence) and nature (homozygous vs heterozygous) of the respective introgression/s.

Data analysis

All statistical data analysis was performed in R programming software. BC_5F_1 and BC_5F_2 families were tested for deviation from expected Mendelian ratio using chi-square tests. Allele and genotype frequencies were obtained from SNP data for the BC_5F_1 families while for the BC_5F_2 families, these frequencies were obtained from SSR genotyping of the subset of donor segments segregating in respective families. Genomic regions spanning at least 5 Mb and containing 3 or more consecutive SNPs with significant distortion ($p < 0.001$) were defined as segregation distortion regions (SDRs). Genomic regions that completely lacked donor alleles for 3 or more consecutive markers spanning at least 5 Mb were defined as Introgression Devoid Regions (IDRs). These definitions have been modified from the ones described in Jiang et al. (2000) and Waghmare et al. (2016) to represent these regions in terms of physical lengths as the expected segregation ratio in the BC_5F_1 generation precluded the construction of a genetic map. The reference genome sequence, genomic sequences spanning the SDRs and IDRs and the list of cotton genes in these regions were extracted from CottonGen (https://www.cottongen.org/data/genome). Gene ontology (GO) enrichment analysis was carried out on all SDRs and IDRs by using the Enrichment Analysis feature on Cotton Functional Genomics Database (https://www.cottonfgd.org).

RESULTS

Marker distribution and genome coverage

Raw sequence data processing, SNP filtering and post-processing was done separately for the two populations, thus resulting in the retention of different number of total SNP markers. A total of 2542 SNP markers ranging from 19 to 174 per chromosome and averaging one marker per 716 kb (Table 1) was used to characterize the *G. barbadense* population while a total of 3345 SNP markers ranging from 65 to 218 per chromosome and averaging one marker per 536 kb was used to characterize the *G. hirsutum* population. In total, the reported physical length of the tetraploid cotton genome is ~2.5 Gb, out of which 1.9 Gb has been anchored to the 26 chromosomes in Jbrowse CottonGen (Zhang et al. 2015). The 2542 SNPs in the *G. barbadense* populations cover 94.15% (1.82 Gb) of the anchored genome ranging from 82 to 99% for individual chromosomes while for the *G. hirsutum* background, the 3345 SNPs cover 92.64% (1.79 Gb) of the anchored genome ranging from 68 to 99% for individual chromosomes (Table 1).

Genomic and sub genomic distribution of *G. hirsutum* introgression into *G. barbadense*

In all, 2471 (97.21%) of the 2542 loci showed *G. hirsutum* introgression in one or more BC_5F_1 plants. One or more introgressed loci were detected on all 26 chromosomes (Fig. 2). For the 190 BC_5F_1 plants genotyped, there were a total of 617 introgressed chromosomal segments (averaging 3.25 segments per BC_5F_1 plant) ranging in size from 1.64 Mb to 83.55 Mb averaging 23.31 Mb (Table 2). A few chromosomes showed introgression over virtually their entire lengths (Fig. 3). However, some chromosomes contained one or more regions that appeared “resistant” to introgression as shown by absence of *G. hirsutum* alleles on three or more consecutive SNP markers spanning at least 5 Mb. At least 16 such regions localized to 12 chromosomes were devoid of *G. hirsutum* alleles (Fig. 2, Table 3). These chromosomal regions lacking *G. hirsutum* alleles spanned lengths of 0.09 to 14.71 Mb with an average span of 3.11 Mb (Table 3).

At the within population level, the proportion of *G. hirsutum* alleles introgressed into the *G. barbadense* background ranged from 0.12 to 27.65% with an average of 4.35% per individual. The

Table 1. Distribution of markers, anchored genome coverage and average retention of donor alleles in *G. hirsutum* and *G. barbadense* backgrounds.

Chr	***G. hirsutum* background**			***G. barbadense* background**		
	# markers	**% genome covered**	**% donor alleles**	**# markers**	**% genome covered**	**% donor alleles**
1	114	99.01	1.53	105	95.74	5.99
2	178	95.53	6.33	141	96.78	5.4
3	212	97.76	2.67	173	92.14	6.33
4	131	94.61	5.28	100	97.63	4.78
5	145	92.67	6.28	117	96.01	5.33
6	218	97.61	2.55	165	98.02	5.32
7	157	96.06	3.05	130	97.86	4.25
8	192	97.57	5.89	174	90.78	4.95
9	140	91.76	4.03	130	94.78	3.33
10	197	94.89	5.73	150	94.78	6.12
11	169	95.28	3.79	174	99.16	4.13
12	162	91.16	5.35	147	95.47	2.29
13	167	94.58	4.12	126	90.34	5.77
14	97	92.19	6.63	44	85.64	2.67
15	113	93.58	5.83	59	95.81	4.63
16	79	89.38	8.58	19	90.15	2.02
17	78	93.18	4.73	47	82.67	2.68
18	82	89.78	4.33	52	98.82	3.67
19	97	84.08	4.07	74	92.49	3.53
20	85	93.21	8.22	51	95.5	1.82
21	101	90.89	4.37	62	94.55	2.76
22	79	76.74	4.14	51	90.4	1.04
23	95	91.56	6.4	62	91.18	1.47
24	65	68.29	5.94	58	98.17	2.41
25	95	94.16	6.82	69	99.19	3.67
26	97	93.69	2.13	62	84.9	4.13

At subgenome retained *G. hirsutum* alleles at a significantly higher (*p* value = 0.012) rate (4.92%) than the Dt subgenome (2.90%) (Fig. 2). Among the 1832 informative At subgenome loci, 1795 (97.98%) showed introgression. Among the 710 informative Dt subgenome loci, 676 (95.21%) showed introgression. Among the 617 chromosomal segments introgressed in the 190 BC_5F_1 families, 426 (69.04%) were introgressed in the At subgenome and 191 (30.96%) were introgressed in the Dt subgenomes (Table 2).

Genomic and sub genomic distribution of *G. barbadense* introgression into *G. hirsutum*

In all, 3292 (98.41%) of the 3345 loci showed *G. barbadense* introgression in one or more BC_5F_1 plants. One or more introgressed loci were detected on all 26 chromosomes. For the 179 BC_5F_1 plants genotyped, there were a total of 722 introgressed chromosomal segments (averaging 4.03 segments per BC_5F_1 plant) ranging in size from 0.12 to 101.05 Mb and averaging 20.48 Mb (Table 2). While a few chromosomes showed introgression over virtually their entire lengths (Fig. 4), some chromosomes contained regions that resisted introgression (Fig. 2, Table 3). At least 5 such regions localized to 4 chromosomes were devoid of *G. barbadense* alleles. These chromosomal regions spanned lengths of 3.92 to 56.75 Mb with an average span of 14.42 Mb (Table 3).

At the within population level, the proportion of *G. barbadense* alleles introgressed into the *G. hirsutum* background ranged from 0.09 to 33.45% with an average of 4.79% per individual. Introgression of *G. hirsutum* chromatin into *G. barbadense* occurred at similar rates (*p* value = 0.91) in the At and the Dt subgenomes (Fig. 2). Unlike the reciprocal population, the At subgenome retained *G. hirsutum* alleles at a slightly lower rate (4.46%) than the Dt subgenome (5.52%). Among the 2182 informative At subgenome loci, 2131 (97.66%) showed introgression. Among the 1163 informative Dt subgenome loci, 1161 (99.82%) showed introgression. Among the 722 chromosomal segments introgressed in the 190 BC_5F_1 families, 355 (49.17%) were introgressed in the At subgenome and 367 (50.83%) were introgressed in the Dt subgenomes (Table 2).

Further, we compared the nature and pattern of introgression of donor chromatin between the centromeric/pericentromeric regions and the telomeric regions of the chromosomes (Fig. 2). The location of the centromeres was obtained from (Hu et al. 2019) based on the TM-1 version-1 genome assembly (the same version used for data analysis). In most cases, the introgression frequency is constant for a long stretch in the centromeric and pericentromeric regions, suggesting that these regions did not experience a lot of meiotic recombination events. In contrast, most telomeric ends have variable introgression frequencies suggesting more and frequent crossover events occurring in the distal regions of the chromosomes.

Segregation distortion and segregation distorted regions (SDRs)

Ideally a BC_5F_1 population is expected to segregate in a 31:1 ratio. A total of 793 markers in the *G. hirsutum* background and 488 in the *G. barbadense* background significantly deviated (χ^2 test, $P < 0.01$) from the expected segregation ratio (Fig. S1). Twelve (1.51%, so marginally above the false positive rate) of the distorted loci showed retention towards the recipient parent in the *G. hirsutum* background while

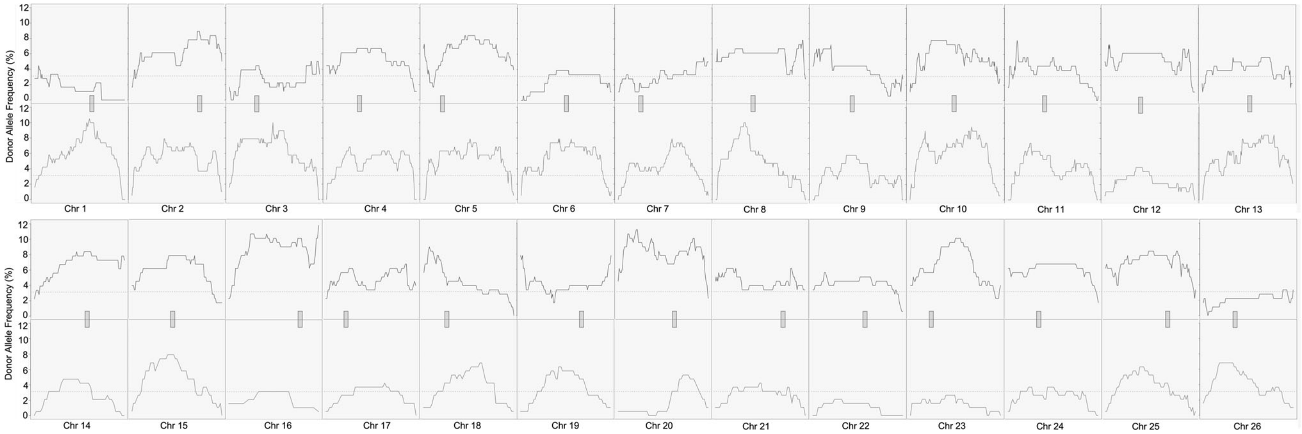

Fig. 2 Retention of donor alleles in reciprocal interspecific populations. X-axis shows markers across the genome separated by chromosomes and y-axis shows the frequency of donor alleles. Rectangular blocks show the tentative location of centromeres across chromosomes. Dotted lines show expected donor frequency (3.125%) for BC_5F_1 generation. Blue lines show donor allele retention in *G. hirsutum* background and red lines show donor allele retention in *G. barbadense* background.

Table 2. Genomic distribution of introgressed chromosomal segments.

	G. barbadense background	*G. hirsutum* background
Total individuals	190	179
Total introgressions	617	722
Average introgressions per plant	3.25	4.03
Min size of int (Mb)	1.67	0.12
Max size of int (Mb)	83.55	101.05
Average size (Mb)	23.31	20.48
At subgenome introgressions	426	355
Dt subgenome introgressions	191	367

781 (98.49%) of the distorted loci retained donor alleles more than expected. In the *G. barbadense* background, all the distorted loci retained donor alleles more than expected. In the *G. hirsutum* background, 462 (58.26%) distorted markers originated from the At subgenome and 331 (41.74%) from the Dt subgenome while in the *G. barbadense* background, 242 distorted markers originated from the At subgenome and 246 originated from the Dt subgenome. A total of 32 SDRs were identified in the two backgrounds (13 in the *G. hirsutum* background and 19 in the *G. barbadense* background) (Table S2). These regions show biased distribution along the length of the chromosomes, with most in the pericentromeric regions and only four (12.5%) near the telomeric regions.

Regions with prominent introgression

Among the 32 SDRs identified in both backgrounds (at $p < 0.01$), five regions in *G. barbadense* background and eight in *G. hirsutum* were significant even at a very stringent statistical measure ($p < 0.0001$). Under further scrutiny, these regions were found to harbor donor alleles in two- to five times the number of individuals than would be expected. The highly introgressed regions, referred to as "regions of prominent introgression" by Wang et al. (1995), were all in the At subgenome in *G. barbadense* background while in the *G. hirsutum* background, 6 (of 8) of these regions were found in the Dt subgenome (Table S2). Of the 13 regions of prominent introgressions identified in the study, 12 were pericentromeric and only one (SDRGh18.1) was near-telomeric. Among the five *G. hirsutum* allele rich regions in the *G. barbadense* background, SDRGh1.1 was also identified in the study by Wang et al. (1995), where the authors identified *G. hirsutum* chromatin in a collection of 54 Sea Island, Egyptian and Pima cottons (*G. barbadense*). The sequence of the RFLP marker A1097 delineating the *G. hirsutum* allele rich region showed DNA sequence correspondence to the same cotton reference genome sequence used in this study. The location of this marker (72708857–72708647 bp, Chr 1, *G. hirsutum* acc TM-1 NAU-NBI genome assembly) was found within the boundaries of SDRGh1.1 (33.5 Mb to 76.8 Mb, Table S2).

Introgression devoid regions (IDRs) and gene ontology enrichment analysis in IDRs

Deviation from expected donor allele frequencies is one of the important features studied in transmission genetics. While significant number of markers spanning several genomic regions (may) deviate from expected frequencies, some genomic regions are totally devoid of donor alleles. While occasional genotyping (sequencing) errors can account for occasional anomalous DNA marker loci, 'runs' of consecutive markers in the genome that are all devoid of introgression cannot realistically be attributed to chance. Regions where 3 or more consecutive markers lack donor alleles were defined as introgression devoid regions (IDRs). A total of 16 IDRs distributed over 12 chromosomes were identified in the *G. barbadense* background while 5 IDRs distributed over four chromosomes were identified in *G. hirsutum* (Table 3).

As genomic regions devoid of donor alleles might harbor genes that are biologically significant for the recipient genome, we looked for genes enriched in these IDRs. A total of 1593 genes were identified in the *G. hirsutum* background, of which 298 belonged to GO terms. Two significantly enriched GO terms were identified, one in chromosome 6 and the other in chromosome 11 (Table 4). Both GO terms were involved in molecular functions related to oxidoreductase and fatty-acid binding activity. In the *G. barbadense* background, a total of 3656 genes were identified, of which 721 belonged to GO terms. Ten significantly enriched GO terms were identified (Table 4), of which three were involved in biological functions (cellulose biosynthesis, recognition of pollen), six in molecular functions and one in cellular functions.

Segregation of donor chromatin in BC_5F_2 families

A total of 190 BC_5F_2 families comprising 2973 individuals (ranging from 2 to 32 and averaging 15.64 individuals per family) in the *G. barbadense* background and 179 BC_5F_2 families comprising 2342 individuals (ranging from 2 to 32 and averaging 13.15 individuals per family) in *G. hirsutum* background were subjected to study of

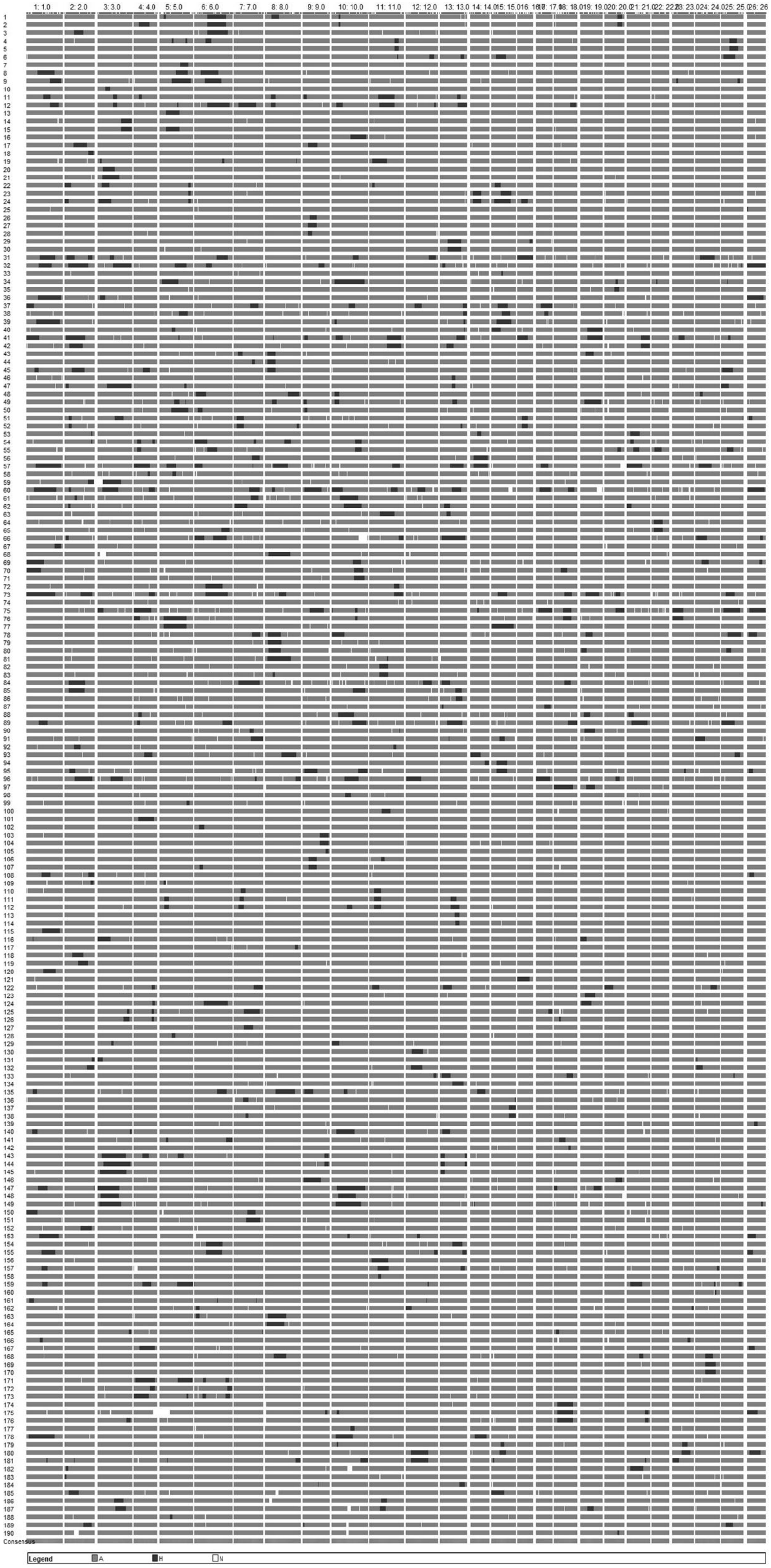

Fig. 3 Genomic composition of 190 BC5F1 lines in *G. barbadense* background. Gray areas represent *G. barbadense* homozygotes (BB), black areas represent heterozygotes (BH) and while areas represent missing genotypes. Chromosomes are shown in the x-axis and individuals are shown in the y-axis.

Table 3. Distribution of introgression devoid regions (IDRs) in *G. hirsutum* and *G. barbadense* backgrounds.

Background	Chr	IDR id	Location (Mb)	# Genes	# GO terms	GO Enriched
	1	IDRGh01.1	93.02–99.77	873	0	0
	3	IDRGh03.1	5.03–08.29	116	0	0
G. hirsutum	6	IDRGh06.1	0.30–04.48	269	178	1
	6	IDRGh06.2	5.86–09.86	134	0	0
	11	IDRGh11.1	87.81–91.73	201	120	1
	1	IDRGb01.1	91.87–96.53	198	0	0
	3	IDRGb03.1	93.61–96.23	131	87	2
	4	IDRGb04.1	56.67–61.88	285	191	1
	5	IDRGb05.1	2.82–04.61	169	0	0
	8	IDRGb08.1	3.12–06.96	228	164	2
	8	IDRGb08.2	97.10–97.19	3	0	0
	9	IDRGb09.1	1.68–02.53	36	0	0
G. barbadense	9	IDRGb09.2	71.51–72.76	150	101	3
	11	IDRGb11.1	0.22–04.59	467	0	0
	11	IDRGb11.2	91.63–92.76	54	0	0
	14	IDRGb14.1	55.75–56.62	23	0	0
	20	IDRGb20.1	24.48–26.51	1072	0	0
	22	IDRGb22.1	0.28–04.75	299	178	2
	22	IDRGb22.2	31.66–46.38	449	0	0
	23	IDRGb23.1	53.83–55.01	0	0	0
	24	IDRGb24.1	64.86–65.59	92	0	0

segregation ratios. At the subset of loci retaining the donor allele in BC_5F_1 plants, segregation ratios observed in the BC_5F_2 progeny showed bias against donor chromatin in both backgrounds. Across all *G. barbadense* chromosomal segments introgressed into *G. hirsutum* BC_5F_1 plants, the average frequency of *G. barbadense* allele retention was 35.42%, much less than the expected 50% (p value < 0.0001). At codominant marker loci, heterozygotes occurred at an average frequency of 32.54 % (versus 50% expected), whereas *G. barbadense* homozygotes occurred at 18.58% (versus 25% expected). At dominant marker loci, 30.61% of individuals had at least one copy of the *G. barbadense* allele (versus 75% expected). In the reciprocal cross, across all *G. hirsutum* chromosomal segments introgressed into *G. barbadense* BC_5F_1 plants, the average frequency of *G. hirsutum* allele retention was 25.85%, much less than the expected 50% ($\chi^2_{2\ \mathrm{df}} = 2959.14$, p value < 0.0001). At codominant marker loci, heterozygotes occurred at an average frequency of 28.49%, whereas *G. hirsutum* homozygotes occurred at 12.32%. At dominant marker loci, 35.02% of individuals had at least one copy of the *G. hirsutum* allele.

Significant deviation from expected genotypic and allelic frequencies was observed for several loci tested in the BC_5F_2 families. In the *G. hirsutum* background, 63.59% of DNA markers distorted significantly from the expected 1:2:1 ratio across individual families ($p < 0.01$) while only 22.82% of the markers were significantly distorted in at least one family for allelic segregation (1:1). Significant distortion from expected genotypic frequency was observed in all chromosomes except 14 and 24 (p value < 0.01) while significant deviation from expected allelic frequencies were observed in all but three chromosomes (14, 18 and 24) (Table S1). Similarly, in the *G. barbadense* background, 76.79% of the markers distorted significantly from the expected 1:2:1 ratio across individual families ($p < 0.01$) while only 39.32% of the markers were significantly distorted in at least one family for allelic segregation (1:1). Significant distortion from expected genotypic frequency was observed in all chromosomes except 17 and 24, while significant deviation from expected allelic frequencies were observed in all chromosomes except 14, 17 and 24 (Table S1).

Individual loci showed significant differences in segregation patterns in different BC_5F_2 families. A total of six loci segregating in three or more families in *G. hirsutum* background are shown in Table 5. Locus DPL0085 is exemplary, showing donor allele retention of 80.95% in family 9037 and 73.08% in family 9103 but only 13.51% in family 9127. Locus CIR0185 shows similar pattern (with 88% donor allele retention). However, for most of the other alleles shown here (and not shown because of retention in less than three families), donor allele retention is significantly lower than expected. In the *G. barbadense* background, a total of eight loci segregating in three or more families in *G. barbadense* background are shown in Table 6. For all eight loci, donor allele retention was significantly lower than expected segregation at the genotypic or allelic level.

Selection against donor alleles in BC_5F_2 families

A total of 13 loci showed no *G. hirsutum* (HH) homozygotes in 181 cases across all the segregating families in the *G. barbadense* background, and a total of 7 loci showed no *G. barbadense* (BB) homozygotes in 205 cases across all the segregating families in *G. hirsutum* background. This suggests a mild level of negative selection against donor alleles at or near these loci. Selection against *G. barbadense* homozygotes was nominally stronger ($\chi^2 = 3.57$, p value = 0.058) at At (85.71%) than Dt subgenomic loci (14.29%) in *G. hirsutum* background. Similarly, in the *G. barbadense* background, selection against *G. hirsutum* homozygotes was nominally stronger ($\chi^2 = 3.53$, p value = 0.061) at At (76.92%) than Dt subgenomic loci (23.08%). Segregation distortion as reflected by genotypic versus allelic frequency ratios indicates mild negative selection against the donor homozygotes. In both backgrounds, genotypic distortion was higher than allelic distortion: in *G. hirsutum* background, 63.59% of markers deviated significantly from genotypic expectations while only 22.82% deviated from allelic expectations; and in *G. barbadense* background, 76.79% markers deviated significantly from genotypic expectations and only 39.32% deviated from allelic expectations.

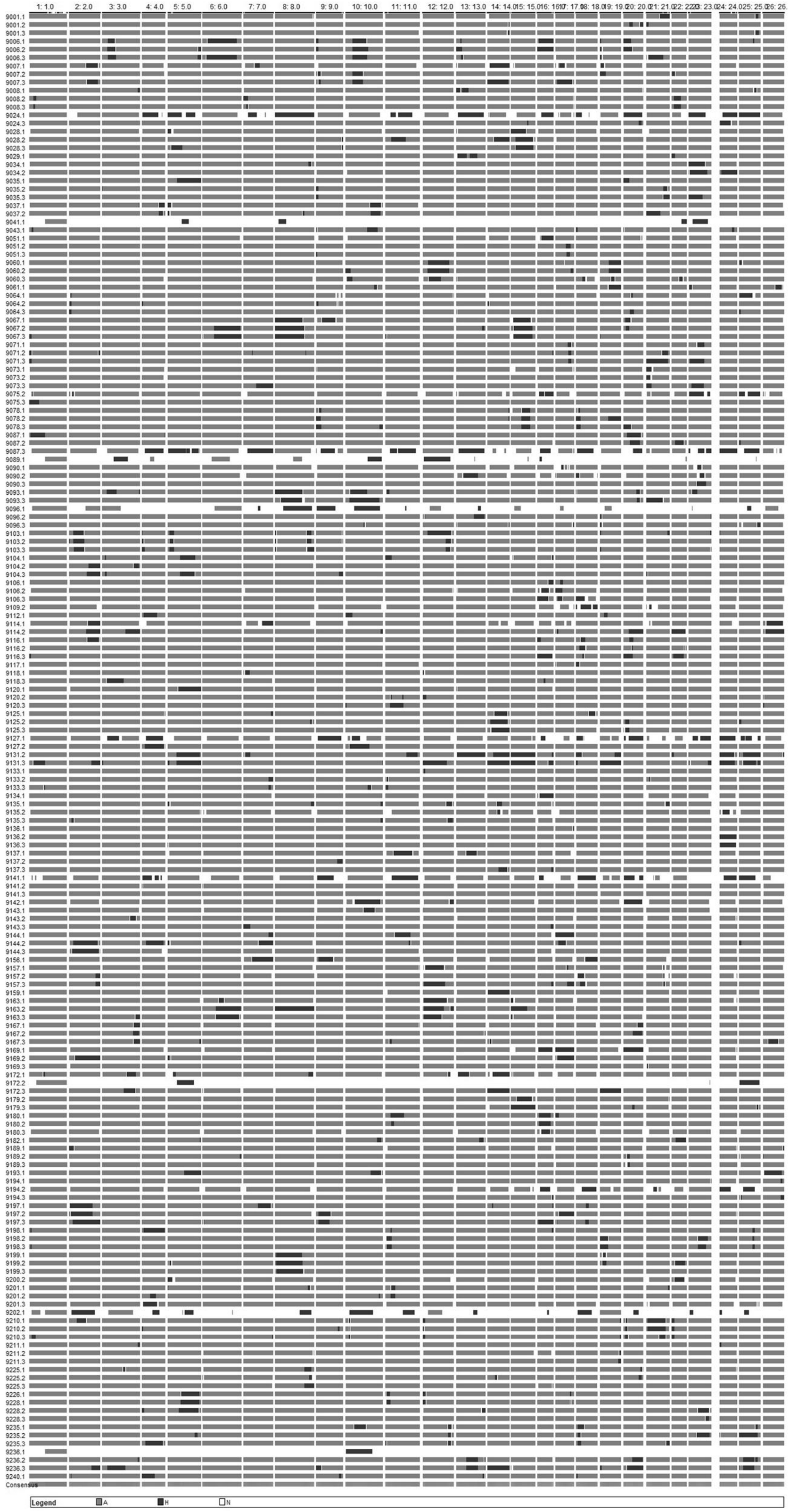

Fig. 4 Genomic composition of 179 BC5F1 lines in *G. hirsutum* background. Gray areas represent *G. hirsutum* homozygotes (HH), black areas represent heterozygotes (HB) and while areas represent missing genotypes. Chromosomes are shown in the x-axis and individuals are shown in the y-axis.

DISCUSSION

A handful of studies has been carried out on transmission genetics of tetraploid cotton species, most of them focusing on the nature and patterns of introgression of donor alleles to the *G. hirsutum* background from *G. barbadense* (Jiang et al. 2000; Stephens 1949), *G. tomentosum* (Waghmare et al. 2016) and *G. mustelinum*

Table 4. GO enriched terms identified in IDRs in both backgrounds.

IDR_ID	Accession	Name	GO Type	*Q* value
IDRGh06.1	GO:0016705	oxidoreductase activity	molecular	9.00E–05
IDRGh11.1	GO:0000062	fatty-acyl-CoA binding	molecular	6.80E–06
IDRGb03.1	GO:0015936	coenzyme A metabolic process	biological	4.70E–07
	GO:0004420	NADPH activity	molecular	4.70E–07
IDRGb04.1	GO:0004506	squalene monooxygenase activity	molecular	1.30E–06
IDRGb08.1	GO:0016853	isomerase activity	molecular	1.40E–05
	GO:0009507	chloroplast	cellular	4.00E–05
IDRGb09.2	GO:0004869	endopeptidase inhibitor activity	molecular	1.00E–05
	GO:0016760	cellulose synthase activity	molecular	4.80E–05
	GO:0030244	cellulose biosynthetic process	biological	8.80E–05
IDRGb22.1	GO:0048544	recognition of pollen	biological	9.00E–07
	GO:0004674	protein serine/threonine kinase activity	molecular	1.00E–05

(Chandnani et al. 2017). The present study extends our knowledge of the patterns of introgression from *G. barbadense* to *G. hirsutum* while also providing novel insights on the nature and patterns of introgression of *G. hirsutum* chromatin into *G. barbadense* in a reciprocal experimental population. Our experimental data provide a glimpse into the consequences of natural exchange of chromatin between these two species. More generally, we provide rich empirical data useful to investigate many issues related to levels and patterns of introgression among species.

Cross compatibility between *G. hirsutum* and *G. barbadense*

These two polyploid species of cotton are cross compatible, but our observations add evidence to the finding that their interspecific hybrids exhibit genetic breakdown during segregation (Hu et al. 2019; Zhang et al. 2014). In the two reciprocal backgrounds, our backcrossing scheme started with ~300–400 F_1 hybrids which declined to ~180–190 BC_5F_1 lines, with each generation of backcrossing seeing loss of some progenies despite our efforts to maintain plants and seeds for each line. Most lines that were lost during generation advancement did not produce viable seeds (evident from low/zero rates of germination) suggesting ongoing genetic breakdown during cycles of crossing. In addition, the amount of abscission observed in crossed flowers (2–5 days after pollination) was also considerably higher than that observed in self-pollinated flowers suggesting that although these two polyploids are cross compatible, they sustain only a fraction of cross-pollinated seeds.

Genomic composition of the BC_5F_1 plants

Across the entire genome, the average retention of *G. hirsutum* alleles at the 2542 assayed loci among the 190 BC_5F_1 families was 4.35%, slightly but not significantly higher than the expected 3.125% ($z = 0.97$ and p value $= 0.165$). Individual loci retained from 0.53 to 10.53% of *G. hirsutum* alleles (Fig. 2) while individual BC_5F_1 families retained *G. hirsutum* alleles at 0.12–27.65%. A total of 48.9% of individuals retained *G. hirsutum* alleles at a rate lower than expected while 51.1% of individuals retained alleles at higher rate.

Similarly, the average retention of *G. barbadense* alleles at 3345 assayed loci among the 179 BC_5F_1 families was 4.79%, slightly but not significantly higher than the expected 3.125% ($z = 1.277$ and p value $= 0.101$). Individual loci retained from 0.55 to 11% of *G. barbadense* alleles while individual BC_5F_1 families retained *G. barbadense* alleles at 0.08–33.45% (Fig. 2). A total of 46.6% of individuals retained *G. barbadense* alleles at a rate lower than expected while 53.4% of individuals retained alleles at higher rate. Average retention of donor alleles was not significantly different for the two reciprocal advanced populations both at the whole-genome level and for individual chromosomes (Table 1).

Meiotic recombination that enables introgression is tightly regulated and crossovers do not occur randomly across the chromosomes, being far more abundant in distal than centromeric regions. While our population is not a natural population and selection of only one seed during generation advancement might truncate information on possible meiotic recombination, we generally observed long stretches around the centromere to have limited crossover events resulting in constant low introgression frequencies.

Persistence of donor chromatin in recipient genome

This study reveals consequences of reciprocal introgression between elite cultivars Acala Maxxa (*G. hirsutum*) and Pima S6 (*G. barbadense*). Higher than expected average levels of introgression of donor chromatin in both backgrounds (*G. hirsutum* background = 4.79 %, *G. barbadense* background = 4.35%) suggest favorability of donor alleles in general. Higher fitness of heterozygotes over homozygous genotypes might be a major cause of these results. Most previous studies showed unintentional selection against donor chromatin in interspecific crosses (Chandnani et al. 2017; Jiang et al. 2000; Waghmare et al. 2016). Small population sizes and lack of genome-wide genetic markers might have caused previous studies to underestimate the level of introgression. A previous study with population size similar to this experiment also reported twice as many distorted loci favoring heterozygous than homozygote state of the recurrent parent (Yu et al. 2011). Although we have used a high density of markers to scan the genome, still 0.2 Gb of genetically anchored genome was lacking polymorphic or segregating markers. One reason for the lack of polymorphism in this proportion of the genome might be the history of introgression from *G. barbadense* to *G. hirsutum* background for Acala cultivar development (Wang et al. 1995). Lack of genetic markers also might reflect uneven genome sampling in GBS libraries.

In the self-pollinated progeny of the BC_5F_1 plants, there was a conspicuous deficiency of donor alleles in both backgrounds. Patterns of segregation in the BC_5F_2 families were similar to those found in previous studies of *G. barbadense*, *G. tomentosum* and *G. mustelinum*, with frequencies of donor alleles that are lower than expected across most segregating families (Chandnani et al. 2017; Jiang et al. 2000; Waghmare et al. 2016). These results suggest that segregation in BC_5F_2 families favor the recipient haplotype with a higher average frequency of the recipient homozygotes than the donor homozygotes, which adds to prior evidence of non-random maintenance of integrity of the recipient genome and further supports the notion that higher fitness of heterozygotes than homozygotes may contribute to persistence of donor chromatin in recipient (recurrent) parent genomes.

Table 5. Loci showing significant variation in segregation patterns as tested by χ^2 test (at $p < 0.01$) in BC_5F_2 families (*G. hirsutum* background).

BC_5F_2	BNL3441			BNL3580			CIR0185		
Family	**HH**	**HB**	**BB**	**HH**	**HB**	**BB**	**HH**	**HB**	**BB**
9006	15	25	17	–	–	–	3	20	12
9035	–	–	–	–	–	–	12	2	1
9037	–	–	–	–	–	–	–	–	–
9075	–	–	–	4	4	0	–	–	–
9087	–	–	–	23	4	1	–	–	–
9093	12	5	3	–	–	–	–	–	–
9103	–	–	–	–	–	–	–	–	–
9118	16	1	0	–	–	–	–	–	–
9120	–	–	–	–	–	–	9	1	9
9127	13	5	7	–	–	–	–	–	–
9131	–	–	–	–	–	–	20	18	6
9157	–	–	–	–	–	–	–	–	–
9167	–	–	–	–	–	–	20	3	1
9198	–	–	–	11	11	9	–	–	–
9201	–	–	–	–	–	–	–	–	–
BC_5F_2	**DPL0085**			**DPL0176**			**MUSB1307**		
Family	**HH**	**HB**	**BB**	**HH**	**HB**	**BB**	**HH**	**HB**	**BB**
9006	–	–	–	–	–	–	–	–	–
9035	–	–	–	–	–	–	–	–	–
9037	6	3	24	–	–	–	–	–	–
9075	–	–	–	–	–	–	–	–	–
9087	–	–	–	-	–	–	–	–	–
9093	–	–	–	–	–	–	–	–	–
9103	7	2	18	10	24	3	8	3	8
9118	–	–	–	–	–	–	–	–	–
9120	–	–	–	–	–	–	–	–	–
9127	32	6	5	–	–	–	–	–	–
9131	–	–	–	–	–	–	13	5	9
9157	–	–	–	–	–	–	3	1	4
9167	–	–	–	–	–	–	–	–	–
9198	–	–	–	–	–	–	–	--	–
9201	–	–	–	1	1	2	–	–	–

Introgression devoid regions

Some regions in the recipient genome are not as tolerant of donor chromatin as other genomic regions. We found a total of 5 regions in the *G. hirsutum* background completely lacking *G. barbadense* introgression, accounting for 1.16% (22.10 Mb) of the total physical length of the anchored cotton genome (Table 3). Curiously, these do not correspond to 7 regions completely lacking *G. barbadense* introgression in a prior study (Jiang et al. 2000), suggesting that even among different combinations of *G. barbadense* and *G. hirsutum*, different chromosomal regions may be devoid of introgression. In the *G. barbadense* background, there were 16 regions completely devoid of *G. hirsutum* introgression, accounting for 2.62% (49.8 Mb) of the anchored genome. This indicates that although each BC_5F_1 individual was introgressed with donor alleles at a slightly higher than expected fraction of loci, introgression was possible in certain regions only. Similar results were reported in studies of introgression of chromatin of wild cotton relatives into *G. hirsutum* (Chandnani et al. 2017; Waghmare et al. 2016) and in the study of introgression of *G. hirsutum* chromatin into *G. barbadense* (Wang et al. 1995).

Segregation distortion was generally evident from multiple linked markers, thus was clearly not attributable to sequencing errors but was a result of biological factors. If the elimination of donor chromatin were to occur randomly after a backcross, then the probability of any one unlinked region lacking introgression in the BC_5F_1 line would be (1–0.03125) = 0.96875. With the simplifying assumption that each of the 5 unlinked regions (in the *G. hirsutum* background) of segregation distortion behaves as a single unit of inheritance and all segregate independently, the probability of all 4 unlinked regions lacking introgressions in all 179 BC_5F_1 plants would be $[(0.96875)^5]^{179} = 4.56 \times 10^{-13}$. Thus, it is unlikely that all these regions lack introgression in all 179 BC_5F_1 plants by chance. In the *G. barbadense* background where there were 16 regions completely devoid of *G. hirsutum* introgression, this probability is $[(0.96875)^{16}]^{190} = 1.21 \times 10^{-42}$, providing even stronger evidence that this lack of introgression is not by chance but due to some biological factors.

To investigate possible biological factors involved in some genomic regions being recalcitrant to introgression, we studied all 21 genomic regions that were devoid of donor chromatin for gene

Table 6. Loci showing significant variation in segregation patterns as tested by χ^2 test (at $p < 0.01$) in BC_5F_2 families (*G. barbadense* background).

BC_5F_2	**BNL3267**			**BNL3790**			**BNL3903**			**BNL4029**		
Family	**HH**	**HB**	**BB**	**HH**	**HB**	**BB**	**HH**	**HB**	**BB**	**HH**	**HB**	**BB**
10001	–	–	–	–	–	–	29	2	6	–	–	–
10002	–	–	–	–	–	–	–	–	–	43	4	19
10004	–	–	–	–	–	–	–	–	–	–	–	–
10019	30	1	32	–	–	–	–	–	–	–	–	–
10104	–	–	–	30	1	4	–	–	–	–	–	–
10109	–	–	–	–	–	–	–	–	–	–	–	–
10110	–	–	–	–	–	–	–	–	–	20	1	10
10117	–	–	–	–	–	–	–	–	–	–	–	–
10126	–	–	–	–	–	–	–	–	–	–	–	–
10131	–	–	–	–	–	–	–	–	–	35	1	8
10141	–	–	–	–	–	–	–	–	–	–	–	–
10145	–	–	–	–	–	–	–	–	–	–	–	–
10146	–	–	–	38	2	7	–	–	–	–	–	–
10148	–	–	–	–	–	–	27	1	14	–	–	–
10150	–	–	–	–	–	–	–	–	–	–	–	–
10152	–	–	–	54	11	5	–	–	–	–	-–	–
10156	–	–	–	–	–	–	–	–	–	–	–	–
10158	14	12	8	–	–	–	–	–	–	–	–	–
10166	–	–	–	–	–	–	–	–	–	–	–	–
10178	16	1	7	–	–	–	–	_	–	–	–	–
10217	–	–	–	–	–	–	7	2	1	–	–	–
10244	–	–	–	–	–	–	–	–	–	–	–	–
BC_5F_2	**DPL0637**			**DPL0652**			**NAU3207**			**NAU5180**		
Family	**HH**	**HB**	**BB**	**HH**	**HB**	**BB**	**HH**	**HB**	**BB**	**HH**	**HB**	**BB**
10001	–	–	–	–	–	–	–	–	–	–	–	–
10002	25	5	9	–	–	–	–	–	–	–	–	–
10004	12	17	8	–	–	–	–	–	–	–	–	–
10019	–	–	–	–	–	–	–	–	–	–	–-	–
10104	–	–	–	–	–	–	–	–	–	–	–	–
10109	–	–	–	36	8	2	–	–	–	–	–	–
10110	–	–	–	–	–	–	–	–	–	–	–	–
10117	7	17	3	–	–	–	–	–	–	–	–	–
10126	–	–	–	34	6	6	–	–	–	–	–	–
10131	–	–	–	–	–	–	–	–	–	–	–	–
10141	–	–	–	–	–	–	–	–	–	37	13	6
10145	–	–	–	–	–	–	14	2	4	–	–	–
10146	–	–	–	–	–	–	–	–	–	–	–	–
10148	–	–	–	–	–	–	–	–	–	39	2	1
10150	–	–	–	–	–	–	33	6	4	–	–	–
10152	–	–	–	–	–	–	–	–	–	–	–	–
10156	–	–	–	37	4	6	–	–	–	–	–	–
10158	–	–	–	–	–	–	–	–	–	–	–	–
10166	–	–	–	–	–	–	–	–	–	36	20	19
10178	–	–	–	–	–	–	–	–	–	–	–	–
10217	–	–	–	–	–	–	–	–	–	–	–	–
10244	–	–	–	–	–	–	18	11	14	–	–	–

ontology (GO) enrichment. A total of 2 GO enriched terms were identified for the *G. hirsutum* background, both being identified as having molecular functions (Table 4). A total of 10 GO enriched terms were identified in *G. barbadense* background, three related to biological processes (pollen recognition, cellulose biosynthesis, and coenzyme A metabolism) and one to cellular function (chloroplast). All the identified GO terms are basic biological processes suggesting that the selective rejection of certain donor chromosomal regions could be related to species integrity and diversification of closely related species. Nevertheless, the identification of these functional candidates offers testable hypotheses why certain genomic regions exclude alien chromatin over others.

To further investigate the extent and nature of synteny in chromosomal regions of *G. hirsutum* and *G. barbadense* genomes that "resist" introgression, we compared the two genomes. None of the 21 IDRs identified in our study fall into inversions identified between the two genomes by Hu et al. (2019). Four IDRs, IDRGh03.1, IDRGb01.1, IDRGb08.2 and IDRGb22.1, fall into regions that show lower than expected SNP frequencies between the two genomes (named LSPR7, LSPR1, LSPR14 and LSPR28 respectively; Hu et al. 2019). Most IDRs were identified in or near the telomeric regions of the chromosomes, with only four being pericentromeric (one each in chromosomes 4, 14, 20 and 22). Despite being in the active crossover regions of the chromosomes, the fact that these regions appeared recalcitrant to introgression supports our hypothesis of biological significance of these IDRs.

Regions of prominent donor chromatin introgression

While some genomic regions are recalcitrant to donor chromatin, others were more tolerant. In general, both backgrounds allowed a slightly higher frequency of donor alleles than expected. All SDRs identified retained higher frequencies of donor alleles than expected. However, certain genomic regions showed greater richness of donor alleles than others. In the *G. hirsutum* background, eight of 19 SDRs were significantly richer in *G. barbadense* alleles than nominally significant SDRs. To investigate whether the transmission of these chromatin segments from *G. barbadense* to *G. hirsutum* occurs randomly, we compared our findings to those of a study to identify *G. barbadense* chromatin in *G. hirsutum* "Sealand" cultivars developed by the Pee Dee breeding program. Among a total of 22 putative *G. barbadense* chromosome segments in Sealand 542 and Sealand 883 backgrounds (Kumar et al. 2019) and 19 SDRs identified in our study, five regions clearly overlapped and two more were in close proximity (Table S4). These results hint at the possibility that the transfer of certain chromatin regions in interspecific crosses might potentially be related to cellular, molecular, or biological functions and might be informative for crop improvement. Indeed, a total of 13 quantitative trait loci (QTLs) related to six fiber quality traits were identified on the *G. barbadense* introgressed chromosomal segments (Kumar et al. 2019).

Similarly, in *G. barbadense* background, five (of 13) SDRs had enriched *G. hirsutum* chromatin (Table S2). Several *G. barbadense* cultivar groups (Pima, Egyptian and Sea Island) are known to harbor prominent *G. hirsutum* enrichment in five regions, one each in chromosomes 1, 5, 14 and 25, and one in unlinked linkage group U01 Wang et al. (1995). Pima S6, the *G. barbadense* cultivar used in our study, contained *G. hirsutum* chromatin in one G. hirsutum rich region identified in our study, SDRGb01.1 in chromosome 1 (as verified by RFLP marker A1097 within the bounds of SDRGb01.1). Therefore, this 'apparent' SDR may be an artifact of a lack of *G. barbadense* alleles in Pima S6 in the region. However, the other four prominent regions as well as the remaining nominal SDRs identified in our study did not find correspond to *G. hirsutum* rich regions in Pima S6 (Wang et al 1995).

Retention of donor alleles at numbers higher than the expected numbers of loci may reflect fitness consequences in recipient backgrounds; while reduced introgression might be linked with factors such as structural rearrangement, multilocus interaction, species integrity and reproductive isolation. GO and GO enrichment analysis was carried out on all SDRs as retention of donor alleles at higher-than-expected frequencies might reflect fitness consequences in recipient backgrounds. We were especially interested in GO terms related to fitness and adaptation in these regions. A total of 2503 GO terms and 16 GO enriched terms were identified (Tables S2, S3). Among the 16 GO enriched terms identified in these SDRs, 11 were related to biological processes. GO terms related to fitness and adaptation (response to freezing, response to biotic stimulus, defense response, photosynthesis and light reaction and different metabolic processes) were identified in regions rich in donor chromatin (Tables S2, S3) providing a starting point to investigate the hypothesis that genomic regions rich in donor chromatin introgression reflect fitness and adaptation behavior in recipient genome.

Subgenomic differentiation in introgression of donor chromatin

Our data about selection against donor alleles further support the notion that different subgenomes have different evolutionary fates. Selection against At subgenomic loci was slightly stronger than Dt subgenomic loci in the *G. hirsutum* background. Perhaps, this may be related to the observations that D genome has higher expression than A genome and most fiber quality QTLs have been mapped on Dt subgenomic loci in allotetraploid cotton (Flagel and Wendel 2010; Rong et al. 2004). Although the respective progenitor genomes for both subgenomes mostly contain common repertoires of genes, they differ largely in DNA quantities and transposable (repeat) element content; the A subgenome having significantly higher amount of these repeat elements than the D subgenome. In addition, the A subgenome is almost as twice as large as D subgenome in terms of DNA content. Despite these facts, studies have shown more genes with expression bias towards the D subgenome and asymmetrically higher gene loss in A subgenome than in D subgenome (Zhang et al. 2015). Li et al. (2015) showed significantly higher mutation frequency and rate of formation of SNPs within intergenic collinear regions of the D subgenome than in the A subgenome, which is consistent with the observation that disproportionately higher frequency of mutation were observed in Cot-filtered non-coding (CFNC) DNA of the D subgenome than the A subgenome (Rong et al. 2012). Albeit the A subgenome is evolving more rapidly than the D subgenome, more domestication pressure towards selecting higher yield and relaxed selection pressure in the A subgenome (Zhang et al. 2015) might have resulted in more fiber related QTLs being mapped into the D subgenome and in more D subgenome homeologs showing higher expression than their A genome counterparts.

Despite some genomic regions (or loci) showing complete selection against donor homozygotes, others showed different levels of permeability. Such differences in permeability of donor alleles by various regions of the recipient genome may indicate differential levels of fitness for the donor alleles (Rieseberg et al. 1999). Complete absence of recipient homozygotes at a few loci (3 cases each in *G. barbadense* and *G. hirsutum* backgrounds) and/or fixation of donor homozygotes at some loci suggest that a single introgression event can be sufficient to fix the donor allele in a population. At the same time, complete absence of donor homozygotes at other loci suggests that some donor alleles dramatically reduce fitness in the recipient genome. The nature of selection has also been an important aspect of segregation studies (Li et al. 2011). In both backgrounds, more loci deviated from genotypic expectations than from allelic expectation in both BC_5F_1 and BC_5F_2 families. This suggests that zygotic selection may be more important than gametic selection in these populations.

Genetic backgrounds and their effects in transmission of donor alleles

Genetic backgrounds can profoundly affect the introgression of a particular chromosomal regions. Albeit introgression was observed across all chromosomes and the rate of overall as well as chromosome wise introgression was not significantly different in the two populations, transmission of certain genomic locations reveals a contrast in how these two backgrounds appeal each other. A total of 16 introgression devoid regions (IDRs) were identified in *G. barbadense* background while only five were identified in *G. hirsutum* background (Table 3). This clearly suggests that *G. barbadense* offers more resistance to *G. hirsutum* chromatin than the reciprocal; and is supported by the fact that a higher number of SDRs enriched in donor alleles were identified in *G. hirsutum* background than in *G. barbadense* background (Table S2). Occasional crosses between improved forms of *G. hirsutum* and *G. barbadense* have led to a degree of genetic exchange that may have mitigated the resistance of Pima S6 chromatin to *G. hirsutum*. Indeed, most improved genotypes of *G. barbadense* are comprised of 5–10% of *G. hirsutum* chromatin, with about two-thirds of those being clustered at five specific locations (Wang et al. 1995). Efforts to introduce *G. barbadense* traits into *G. hirsutum* cultivars (Kumar et al. 2019) have had much less impact on the elite gene pool, perhaps leaving the inherent isolation mechanisms of *G. hirsutum* more intact.

Other differences between the behavior of introgressed chromatin between these two species are not readily explained by experimental design or breeding history. For example, among those loci that were heterozygous in BC_5F_1 and were segregating in the BC_5F_2 families, both backgrounds showed considerable tolerance of homozygosity from the donor (18.58% homozygosity tolerance by *G. hirsutum* and 12.32% homozygosity tolerance by *G. barbadense*). Similar level of tolerance (18.6%) of *G. barbadense* by *G. hirsutum* was reported by (Waghmare et al 2016). These levels of tolerance were much higher than those reported on a wide cross involving *G. tomentosum* (1.27%) in *G. hirsutum* background (Waghmare et al. 2016). These results are incongruous with the closer geographic proximity of wild *G. hirsutum* (Central America) to *G. barbadense* (Peru) that would seem to confer a greater selective advantage to reproductive-isolation mechanisms acting between these species than with *G. tomentosum* (Hawaii). The greater evolutionary distance between *G. hirsutum* and *G. barbadense* (representing different polyploid clades) should also have provided greater opportunity for such mechanisms to evolve.

Reciprocal transmission of donor chromatin

The reciprocal populations described here offer a broader scope of understanding the genetics of transmission of donor chromatin than can be achieved by more conventional, unidirectional, studies. The overall rate of introgression of donor chromatin in the reciprocal populations was similar (4.22 in *G. barbadense* background vs 4.79% in *G. hirsutum* background), but the nature of this retention was very different when we look closely at specific genomic regions (Fig. 2, Table 1). For example, in the *G. barbadense* background, the retention of donor alleles along the length of chromosome 1 (5.99% for this chromosome) is almost always greater than the expected rate of 3.125% suggesting favorability of *G. hirsutum* chromatin along the length of chromosome 1. Interestingly, in the reciprocal (*G. hirsutum*) background, the donor (*G. barbadense*) alleles are retained at rates lower than expected along the length of chromosome 1, suggesting selection favoring *G. hirsutum* alleles. This hypothesis is further supported by the fact that no *G. barbadense* introgressions were identified in chromosome 1 in two crosses involving upland cotton Suyuan 7235 as female parent and Sealand 542 and Sealand 883 as male parents respectively in just two generations after initial crossing (Kumar et al 2019). In addition, prominent regions of *G. hirsutum* chromatin introgression was observed in a large and wide collection of *G. barbadense* cultivars (Wang et al. 1995) with some Pima and Sea Island accessions harboring *G. hirsutum* chromatin along the entire length of chromosome 1. Recent deep sequencing study conducted by Hu et al. (2019) in nine *G. barbadense* and ten *G. hirsutum* accessions also revealed an *G. hirsutum* chromatin introgressed region (43.10 Mb to 92.00 Mb) on chromosome 1 in all nine *G. barbadense* accessions collected from Egyptian, American Pima and Central Asian ELS cottons.

Selection favoring *G. barbadense* alleles is exemplified by chromosome 23. The *G. hirsutum* background retained *G. barbadense* alleles at 6.4% for chromosome 23 while the reciprocal background harbored *G. hirsutum* alleles at significantly lower rates (p value < 0.0001) than expected. Indeed, chromosomes 3, 6, 9, 12, 13, 16, 17, 18, 19, 20 and 26 contain short regions favoring chromatin from one species over the other. Certain regions, however, have shown heterozygote advantage over the recipient alleles, retaining donor alleles in higher frequencies than expected in both backgrounds (Figs. 2 and 3). For example, almost the entire length of chromosome 2 shows favorability for donor alleles in both backgrounds (6.33% retention in *G. hirsutum* background and 5.4% retention in *G. barbadense* background). Similar patterns of heterozygote advantage have been observed in regions of chromosomes 4, 5, 10, 15 and 25. Other genomic regions completely alienated donor chromatin in both backgrounds. Although almost the entire length of chromosome 1 favored *G. hirsutum* alleles, the distal end of both this chromosome and chr. 11 lacked donor alleles in both reciprocal crosses (Figs. 2 and 3, Table 3). Avoidance of donor chromatin may be related to species integrity via preservation of important cellular, molecular, and/or biological functions.

CONCLUSION

In summary, reciprocal transmission genetic study between *G. hirsutum* and *G. barbadense* shows that the extent of introgression and the fate of introgressed chromatin depends on several factors including genetic background, fitness of substituted alleles and allelic combinations, and location of transmitted chromatin. An important motivation for the analysis of advanced-generation interspecific population involving these two species is that they have been frequently analyzed for discovery of novel variation that might enhance agricultural productivity. Commercially, *G. barbadense* fiber has qualities superior to those of most if not all *G. hirsutum;* and commands a premium price (currently near 3x!) though is much lower yielding. Valuable phenotypic attributes associated with *G. barbadense* introgression have been reported (Jiang et al. 2000). The same population also revealed a rich set of QTLs with potentially desirable attributes (Chee et al. 2005; Draye et al 2005), a subset of which have been studied in detail for their value in elite germplasm (Shen et al. 2011). The transfer of desirable attributes to/from *G. hirsutum* to/from *G. barbadense*, long a goal of many cotton breeders, has generally failed. Genetic analysis now provides insight into the biological complexity of this undertaking, which is complicated by interactions between unlinked loci, pronounced differences among genetic backgrounds, uncertain predictive value across generations, and difficulties associated with obtaining fixed (homozygous) genotypes for many introgressed segments.

DATA AVAILABILITY

The raw data supporting the conclusions of this article has been included in the supplementary information (Dataset 1 and Dataset 2) and has also been deposited in the Dryad repository (https://doi.org/10.5061/dryad.kh189329v).

REFERENCES

Adhikari J, Das S, Wang Z, Khanal S, Chandnani R, Patel JD et al. (2017) Targeted identification of association between cotton fiber quality traits and microsatellite markers. Euphytica 213(3):65

Anderson E (1949). Introgressive hybridization. J. Wiley

Andolfatto P, Davison D, Erezyilmaz D, Hu TT, Mast J, Sunayama-Morita T et al. (2011). Vol. 21. COLD SPRING HARBOR LABORATORY PRESS: United States, pp 610-617

Baack E, Melo MC, Rieseberg LH, Ortiz-Barrientos D (2015) The origins of reproductive isolation in plants. N Phytol 207(4):968–984

Beasley JO (1942) Meiotic chromosome behaviour in species, species hybrids, haploids, and induced polyploids of Gossypium. Genetics 27:25–54

Brubaker CL, Paterson AH, Wendel JF (1999) Comparative genetic mapping of allotetraploid cotton and its diploid progenitors. Genome 42(2):184–203

Chandnani R, Wang B, Draye X, Rainville LK, Auckland S, Zhuang Z et al. (2017) Segregation distortion and genome-wide digenic interactions affect transmission of introgressed chromatin from wild cotton species. Theor Appl Genet 130(10):2219

Chee PW, Draye X, Jiang CX, Decanini L, Delmonte TA, Bredhauer R et al. (2005) Molecular dissection of phenotypic variation between Gossypium hirsutum and Gossypium barbadense (cotton) by a backcross-self approach: III. Fiber length. TAG Theor Appl Genet 111(4):772–781

Desai A, Chee PW, Junkang R, May OL, Paterson AH, Gustafson JP (2006) Chromosome structural changes in diploid and tetraploid A genomes of Gossypium. Genome 49(4):336–345

Draye X, Chee P, Jiang C-X, Decanini L, Delmonte TA, Bredhauer R et al. (2005) Molecular dissection of interspecific variation between Gossypium hirsutum and G. barbadense (cotton) by a backcross-self approach: II. Fiber fineness. Theor Appl Genet 111(4):764–771

Flagel LE, Wendel JF (2010) Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. N Phytol 186(1):184–193

Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q et al. (2014) TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. PLoS ONE 9(2):1–11

Grant V (1981). Plant speciation, 2nd edn. Columbia University Press

Hajjar R, Hodgkin T (2007) The use of wild relatives in crop improvement: a survey of developments over the last 20 years. Euphytica 156(1):1–13

Heiser CB (1979) Hybrid Populations of Helianthus divaricatus and H. microcephalus after 22 Years. Taxon 28(1/3):71–75

Hu Y, Chen J, Fang L, Zhang Z, Ma W, Niu Y et al. (2019) Gossypium barbadense and Gossypium hirsutum genomes provide insights into the origin and evolution of allotetraploid cotton. Nat Genet 51(4):739–748

Jiang C-X, Chee PW, Draye X, Morrell PL, Smith CW, Paterson AH (2000) Multilocus interactions restrict gene introgression in interspecific populations of polyploid gossypium (cotton). Evolution 54(3):798–814

Kantartzi S, Roupakias DG (2008) Breeding barriers between Gossypium spp. and species of the Malvaceae family. Aust J Bot 56(3):241–245

Kelly S, Huihui L, Navarro JAR, Dong A, Maria Cinta R, Sarah H et al (2014). Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. The Plant Genome 7(3):1–12

Kim C, Guo H, Kong W, Rahul C, Shuang L, Paterson AH (2016) Application of genotyping by sequencing technology to a variety of crop breeding programs. Plant Sci 242:14–22

Kumar P, Singh R, Lubbers EL, Shen X, Paterson AH, Campbell BT et al. (2019) Genetic evaluation of exotic chromatins from two obsolete interspecific introgression lines of upland cotton for fiber quality improvement. Crop Sci 59(3):1073–1084

Levi A, Ovnat L, Paterson AH, Saranga Y (2009) Photosynthesis of cotton near-isogenic lines introgressed with QTLs for productivity and drought related traits. Plant Sci 177(2):88–96

Li F, Fan G, Lu C, Xiao G, Zou C, Kohel RJ et al. (2015) Genome sequence of cultivated Upland cotton (Gossypium hirsutum TM-1) provides insights into genome evolution. Nat Biotechnol 33(5):524–530

Meyn O, Emboden WA (1987) Parameters and consequences of introgression in salvia apiana x S. mellifera (Lamiaceae). Syst Bot 12(3):390–399

Paterson AH, Boman RK, Brown SM, Chee PW, Gannaway JR, Gingle AR et al. (2004) Reducing the genetic vulnerability of cotton. Crop Sci 44(6):1900

Paterson AH, Brubaker CL, Wendel JF (1993) A rapid method for extraction of cotton (Gossypium spp.) genomic DNA suitable for RFLP or PCR analysis. Plant Mol Biol Report 11(2):122–127

Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D et al. (2012) Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. Nature 492(7429):423–427

Rieseberg LH, Kim MJ, Seiler GJ (1999) Introgression between the cultivated sunflower and a sympatric wild relative, helianthus petiolaris (Asteraceae). Int J Plant Sci 160(1):102–108

Rong J, Wang X, Schulze SR, Compton RO, Williams-Coplin TD, Goff V et al. (2012) Types, levels and patterns of low-copy DNA sequence divergence, and phylogenetic implications, for Gossypium genome types. Heredity (Edinb) 108(5):500–506

Rong JK, Abbey C, Bowers JE, Brubaker CL, Chang C, Chee PW et al. (2004) A 3347-locus genetic recombination map of sequence-tagged sites reveals features of genome organization, transmission and evolution of cotton (Gossypium). Genetics 166(1):389–417

Shen X, Cao Z, Singh R, Lubbers EL, Xu P, Smith CW et al. (2011) Efficacy of qFL-chr1, a Quantitative Trait Locus for Fiber Length in Cotton (Gossypium spp.). Crop Sci 51(5):2005–2010

Stephens SG (1949) The cytogenetics of speciation in Gossypium I. Selective elimination of the donor parent genotype in interspecific backcrosses. Genetics 1949 34:627–637

Tanksley SD, Nelson JC (1996) Advanced backcross QTL analysis: a method for the simultaneous discovery and transfer of valuable QTLs from unadapted germplasm into elite breeding lines. Theor Appl Genet 92(2):191–203

Waghmare VN, Rong J, Rogers CJ, Bowers JE, Chee PW, Gannaway JR et al. (2016) Comparative transmission genetics of introgressed chromatin in Gossypium (cotton) polyploids. Am J Bot 103(4):719–729

Waghmare VN, Rong J, Rogers CJ, Pierce GJ, Wendel JF, Paterson AH (2005) Genetic mapping of a cross between Gossypium hirsutum (cotton) and the Hawaiian endemic Gossypium tomentosum. Theor Appl Genet 111(4):665–676

Wang GL, Dong JM, Paterson AH (1995) The distribution of Gossypium hirsutum chromatin in G. barbadense germ plasm: molecular analysis of introgressive plant breeding. Theor Appl Genet 91(6):1153–1161

Wendel JF (1989) New world tetraploid cottons contain old world cytoplasm. Proc Natl Acad Sci USA 86(11):4132–4136

Wendel JF, Schnabel A, Seelanan T (1995) Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (Gossypium). Proc Natl Acad Sci USA 92(1):280–284

Yu J, Yu S, Lu C, Wang W, Fan S, Song M et al. (2007) High-density linkage map of cultivated allotetraploid cotton based on SSR, TRAP, SRAP and AFLP markers. J Integr Plant Biol 49(5):716–724

Yu Y, Yuan D, Liang S, Li X, Wang X, Lin Z et al. (2011) Genome structure of cotton revealed by a genome-wide SSR genetic map constructed from a BC1 population between gossypium hirsutum and G. barbadense. BMC Genomics 12(1):15

Zhang J, Stewart J (2004) Semigamy gene is associated with chlorophyll reduction in cotton. Crop Sci 44:2054–2062

Zhang JF, Percy RG, McCarty JC (2014) Introgression genetics and breeding between Upland and Pima cotton: a review. Euphytica 198:1–12

Zhang T, Hu Y, Jiang W, Fang L, Guan X, Chen J et al. (2015) Sequencing of allotetraploid cotton (Gossypium hirsutum L. acc. TM-1) provides a resource for fiber improvement. Nat Biotechnol 33(5):531–537

ACKNOWLEDGEMENTS

We are thankful to the members of Plant Genome Mapping Laboratory for their help in DNA extractions and SSR genotyping.

AUTHOR CONTRIBUTIONS

JA conceived, designed, and performed experiments, developed populations, conducted data analysis and interpretation, and drafted and revised the manuscript. RC performed initial crosses. DV and WP helped in DNA extraction and SSR genotyping, SK performed data analysis and revised manuscript. AHP conceived the project, acquired the funds, supervised the project, and revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

Financial support is acknowledged from the US National Science Foundation (DBI 0817707, AIR 1127755), Cotton Incorporated (14–460), and a Regents Professorship to AHP.

COMPETING INTERESTS

The authors declare no competing interests..

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41437-023-00594-w.

Correspondence and requests for materials should be addressed to Andrew H. Paterson.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.