Insights Into Upland Cotton (*Gossypium hirsutum* L.) Genetic Recombination Based on 3 High-Density Single-Nucleotide Polymorphism and a Consensus Map Developed Independently With Common Parents

Genomics Insights
Volume 10: 1–15

DOI: 10.1177/1178631017735104

Mauricio Ulloa[1], Amanda M Hulse-Kemp[2], Luis M De Santiago[2,3], David M Stelly[2] and John J Burke[1]

[1]USDA-ARS, PA, Plant Stress and Germplasm Development Research Unit, Lubbock, TX, USA.
[2]Department of Soil and Crop Sciences, Texas A&M University, College Station, TX, USA.
[3]Interdisciplinary Program in Genetics, Texas A&M University, College Station, TX, USA.

ABSTRACT: High-density linkage maps are vital to supporting the correct placement of scaffolds and gene sequences on chromosomes and fundamental to contemporary organismal research and scientific approaches to genetic improvement, especially in paleopolyploids with exceptionally complex genomes, eg, upland cotton (*Gossypium hirsutum* L., "2n = 52"). Three independently developed intraspecific upland mapping populations were analyzed to generate 3 high-density genetic linkage single-nucleotide polymorphism (SNP) maps and a consensus map using the CottonSNP63K array. The populations consisted of a previously reported F_2, a recombinant inbred line (RIL), and reciprocal RIL population, from "Phytogen 72" and "Stoneville 474" cultivars. The cluster file provided 7417 genotyped SNP markers, resulting in 26 linkage groups corresponding to the 26 chromosomes (c) of the allotetraploid upland cotton $(AD)_1$ arisen from the merging of 2 genomes ("A" Old World and "D" New World). Patterns of chromosome-specific recombination were largely consistent across mapping populations. The high-density genetic consensus map included 7244 SNP markers that spanned 3538 cM and comprised 3824 SNP bins, of which 1783 and 2041 were in the A_t and D_t subgenomes with 1825 and 1713 cM map lengths, respectively. Subgenome average distances were nearly identical, indicating that subgenomic differences in bin number arose due to the high numbers of SNPs on the D_t subgenome. Examination of expected recombination frequency or crossovers (COs) on the chromosomes within each population of the 2 subgenomes revealed that COs were also not affected by the SNPs or SNP bin number in these subgenomes. Comparative alignment analyses identified historical ancestral A_t-subgenomic translocations of c02 and c03, as well as of c04 and c05. The consensus map SNP sequences aligned with high congruency to the NBI assembly of *Gossypium hirsutum*. However, the genomic comparisons revealed evidence of additional unconfirmed possible duplications, inversions and translocations, and unbalance SNP sequence homology or SNP sequence/loci genomic dominance, or homeolog loci bias of the upland tetraploid A_t and D_t subgenomes. The alignments indicated that 364 SNP-associated previously unintegrated scaffolds can be placed in pseudochromosomes of the NBI *G hirsutum* assembly. This is the first intraspecific SNP genetic linkage consensus map assembled in *G hirsutum* with a core of reproducible mendelian SNP markers assayed on different populations and it provides further knowledge of chromosome arrangement of genic and nongenic SNPs. Together, the consensus map and RIL populations provide a synergistically useful platform for localizing and identifying agronomically important loci for improvement of the cotton crop.

KEYWORDS: Genetic mapping, linkage analysis, recombination, genetics, breeding, molecular markers, SNP, mapping population, recombinant inbred line

RECEIVED: April 16, 2017. **ACCEPTED:** September 10, 2017.

PEER REVIEW: Three peer reviewers contributed to the peer review report. Reviewers' reports totaled 1566 words, excluding any confidential comments to the academic editor.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was funded by USDA-ARS (project 3096-21000-019-00-D) (M.U. and J.J.B.) and Cotton Incorporated (CA State Support Committee), Cary, NC (ARIS log nos. 58-3096-002 and -009 [M.U.] and 58-3096-5-017 [J.J.B.]) and supported by Cotton Incorporated grants under #13-694 and NSF-PGRP grant #1444552 (D.M.S. and A.M.H-K) for the CottonSNP63K array.

DECLARATION OF CONFLICTING INTEREST: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Mauricio Ulloa, USDA-ARS, PA, Plant Stress and Germplasm Development Research Unit, Lubbock, TX 79415, USA.
Email: Mauricio.Ulloa@ars.usda.gov

Introduction

Molecular linkage maps based on DNA markers serve as the backbone for genetic analyses and are widely recognized as an essential tool for genetic research in many species.[1–6] In addition, high-density genetic linkage maps provide an excellent framework for discovering loci and/or genes responsible for traits of interest, and a high-quality map with densely spaced sequence markers is vital to correct placement of scaffolds and gene sequences into chromosomal assemblies.[6] Moreover, high-density linkage maps are fundamental to scientific approaches that will lead to genetic improvement, especially in polyploid and paleopolyploid organisms with exceptionally complex genomes such as upland cotton (*G hirsutum* L., 2n = 52, $2(AD)_1$).

Cotton is the most important renewable natural textile fiber worldwide. Although global cultivation involves 4 *Gossypium* species, only 2, *G hirsutum* and *G barbadense* L., are widely cultivated. Over 95% of US cotton production derives from high-yielding upland cultivars of *G hirsutum* (USDA National Agricultural Statistics Service 2013 http://www.ers.usda.gov). The genus *Gossypium* comprises more than 50 extant species of which at least 45 are regarded as diploid with 2n = 2x = 26 chromosomes and at least 6 are known as allotetraploid with 2n = 4x = 52 chromosomes.[7–10] The common AD genome architecture among all extant 52-chromosome *Gossypium* species is thought to reflect a relatively recent, 1 million years ago monophyletic origin via a

common ancient polyploid that arose between now-extinct species most closely approximated among extant species by *Gossypium arboreum* and *Gossypium herbaceum* (A genome) and *Gossypium raimondii* (D_5 genome).[11] Although both extant A-genome species produce lint (textile fibers) on their seed, none of the extant D-genome species produce lint or commercial fibers.[11,12] Due to the important economic nature and the complex genetic structure of the allotetraploid, there is much intrigue regarding the roles of chromosome biology, hybrid vigor, epigenetics, and transcriptomics in productivity seen in the cultivated tetraploids. These and other interests have further fueled the practical desires to create first-rate genomics resources for *Gossypium* research[13] to guide in building a high-quality genome reference sequence to aid in research and breeding endeavors.

In addition to demonstrating patterns of genomic meiotic affinity and relatedness among 26-chromosome and 52-chromosome species, early cotton researchers observed disomic patterns of inheritance in the early 1900s,[14] and cytogeneticists demonstrated that the 52-chromosome cottons, *G hirsutum* and *G barbadense*, were of AD allotetraploid genome composition and undergo strictly bivalent pairing.[15] Comparative mapping has reinforced the findings, indicating a high level of gross subgenome integrity in the AD genomes. It seems that differences between A_t and D_t subgenomes of the new AD polyploid durably minimized or precluded the occurrence of meiotic interactions (or perpetuation of their products) and afforded considerable A_t-subgenomic and D_t-subgenomic integrity across many generations, at least in surviving lineages. Relative to extant A_1 and A_2 genomes, they demonstrated that the *G hirsutum* A_t subgenome had evolutionarily undergone 2 translocations relative the A_1, and that A_2 had another relative to A_1, too.[16,17] Gross structural changes were not noted for the D_t subgenome. Comparative mapping of the A_t versus D_t subgenomes with homeologous molecular markers and later sequence assemblies subsequently demonstrated the expected gross translocation differences for 2 pairs of A-subgenome chromosomes or linkage groups (LGs) versus the D-subgenome counterparts: c02/c03 versus c14/c17 and c04/c05 versus c19/c22. There are 26 disomic pairing gametic chromosomes, where exchange of genomic regions occurred and/or genetic recombination occurred.[18–22]

A widely recognized essential tool in many crops that serve as the backbone for gametic chromosome paring, recombination, or genetic analyses is a molecular linkage map based on DNA markers.[1,2,4–6] The development of linkage maps or genetic mapping in the last decade was primarily performed with simple sequence repeat (SSR) markers (http://www.cottongen.org).[20,23–25] However, distribution and numbers of SSRs are limited in a genome and have been primarily limited to inclusion of a couple hundred single-nucleotide polymorphism (SNP) markers with SSRs.[20,26,27] The initial SNP marker development in cotton was slow and costly,[28–30] and few SNP markers were made available in the past decade. In addition, initial efforts to develop SNP markers were hindered by the co-identification of SNP interlocus variants between the 2 subgenomes in the tetraploids or homeo-SNPs.

With the availability of next-generation sequencing (NGS) technology, sequencing has become faster and cheaper, recently helping to identify larger number of SNP markers.[26,31–34] Considerable progress has been made toward the development of new cotton genomic resources. The larger collection of SNPs (up to 90 000) was assembled from gene transcripts and genomic DNA of multiple cultivars, genotypes, and species (Cotton—SNP Chip, Illumina BeadArray, Illumina Inc., Mira Loma, CA, USA, and public institutions). Recently, a CottonSNP63K Illumina Infinium array (Illumina Inc.) was validated with 1156 samples, providing more than 7000 upland intraspecific and 19 000 interspecific SNP markers that were amenable to mapping in 2 F_2 populations.[21]

In combination with the recently published cotton ancestor diploid genomes, A genome (*G arboreum*—A_2)[35] and D genome (*G raimondii*—D_5),[36,37] and tetraploid genomes, cultivated upland $(AD)_1$ *G hirsutum* acc. TM-1 genome[38,39] and cultivated Pima $(AD)_2$ *G barbadense* genome[40] genome references, new high-density genetic linkage maps will provide an additional and stronger foundation for fine mapping and genetic dissection of candidate genes and quantitative trait loci (QTL) for important traits such as yield and fiber quality traits, drought and plant stress tolerance, and pest and disease resistance. In addition, mapping multiple populations and developing consensus maps will help to reduce large gaps due to the lack of polymorphism in certain complex genomic regions, to increase the number of mapped loci, to validate marker order, and to increase marker genome coverage.[4,41,42]

The objectives of this study were as follows: (1) to develop genetic linkage maps from 3 independently developed populations and a consensus using the recently developed CottonSNP63K Illumina Infinium array for genetic analysis, (2) to increase our understanding of genetic recombination and genome organization based on developed high-density SNP maps and a consensus map of upland Cotton (*G hirsutum*), and (3) to provide a framework for the discovery of loci/genes of important cotton traits. The 3 intraspecific mapping population SNP maps (a F_2, a F_7 recombinant inbred line [RIL], and F_7 a reciprocal RIL) and a consensus map developed using the CottonSNP63K array advance our understanding and provide additional insights into upland cotton genetic recombination by examining parental relationships, segregation/recombination, and gene/SNP marker order from F_2 population to F_7 generation, and genome organization of the cotton crop. The linkage maps and consensus map will also be used to identify important agronomic, physiological, and fiber quality QTL in the cotton crop.

Materials and Methods

Mapping populations

Two distinct heirloom cultivars from different geographical cotton-growing regions of the United States were used to

independently develop 3 mapping populations with common parents. During their period of cultivation, "Phytogen 72" (PHY72) was grown in the far western region, and "Stoneville 474" (STV474) was grown in the mid-south region. Pollen from these 2 cultivars was collected and transferred to water-treated recipient flowers and fertilized according to the methods of Burke.[43] To develop each of the 3 mapping populations, seed were collected at maturity from a single plant. An F_2 PHY72 × STV474 population with 93 individuals was the first population used in this study.[21] In addition, 2 RIL populations PHY72 × STV474 (132 RILs) and reciprocal STV474 × PHY72 (104 RILs) were also used in different genetic and genomic analyses. Recombinant inbred line populations were developed by single-seed/plant descent from the F_2 to the F_7 generation. Generations were advanced without intentional selection for agronomic, yield or fiber quality traits.

Cotton seed was planted into #5 Custom (20 cm diameter, 7.0 L) pots (BWI Companies, Nash, TX, USA) containing Sunshine Mix #1 soil (Sun Gro Horticulture Distributors Inc., Agawam, MA, USA) at the USDA-ARS, PA, CRSL, Plant Stress and Germplasm Development Research, Lubbock, TX, USA. Three seeds per pot were planted, and pots were placed on benches in a greenhouse set to provide a 31°C/27°C day/night cycle. Plants were thinned to 1 plant per pot and grown throughout the year. Nutrients were maintained by daily application with Peters Excel fertilizer (Scotts-Sierra Horticultural Products Company, Marysville, OH, USA) through an automated watering system.

Genotyping with the CottonSNP63K array

DNA was extracted from young leaves using the Macherey-Nagel Plant NucleoSpin Kit (Macherey-Nagel Inc., Bethlehem, PA, USA) following manufacturer's instructions. All DNA samples were quantified using PicoGreen and then diluted to 50 ng/μL. Standardized DNA at 50 ng/μL was used to genotype with the CottonSNP63K array for all cotton lines at Texas A&M University using the cluster file developed.[21] Briefly, single-base extension was performed to assay SNP positions on the array with fluorescent labels, and the chips were then read using the Illumina iScan. Data image files for all samples were imported into the GenomeStudio software, and genotypes were called with the imported cluster positions from the cluster file using the Genotyping Module (V 1.9.4; Illumina, Inc.).

Genetic linkage analysis

Genotypic SNP data were transformed into "ABH" coding mapping data format for the samples of all mapping populations. In the F_2 population, genotypic matrix data were obtained from,[21] and only markers that behaved codominantly and had opposite homozygous allele calls between parental samples were retained. In the RIL populations, markers were filtered for positions where parents had opposite homozygous alleles (neither parent could be heterozygous). We retained samples with less than 5% missing data, and frequency of either homozygous genotype greater than 5%. Herein, the genotypic matrix data obtained[21] were used to develop a new map with the same software program as the RIL populations to advance our understanding in genetic recombination by examining segregation and gene/SNP marker order from F_2 population to F_7 generation.

For each population, the JoinMap v4.1[44] computer program was used to test the χ^2 goodness of fit for expected versus observed genotypic ratios, and all loci were retained for analysis. The *independence LOD* (logarithm of the odds) was used to develop the grouping node or LGs. The test for independence is not affected by segregation distortion such as the LOD scores normally used in linkage analysis, or *linkage LOD*, providing less spurious linkages. LOD scores of 4 to 16 were examined to determine the final groupings. The maximum likelihood mapping algorithm was used for ordering the markers in each of the LGs with the default grouping setting on the groups/chromosomes in each population and the Kosambi map function. Then, the consensus map program was used to assemble the 3 maps with the *Combine Groups for Map Integration* function. The consensus map was developed using the regression mapping algorithm and the Kosambi map function with a maximum distance of 40 cM.

The final maps were drawn with MapChart version 2.2[45] with one to several markers per centimorgan to allow easier visualization. The correlation between the map orders was also visualized using MapChart. The number of recombination events per individual was calculated for all groups in all populations in JoinMap. After *Expected Recombination Frequency* or number of COs per individual were obtained by the *consensus map* 4.1 program, averages and visualization of CO for each determined LG and joint group were calculated and plotted using Microsoft Excel. In addition, CheckMatrix (http://www.atgc.org/XLinkage/MadMapper/; http://code.google.com/p/atgc-map/) was used to generate heat maps based on recombination scores to assess marker order. One-way analysis of variance (ANOVA) was performed. Fisher Protected Least Significant Difference (LSD) test was also used to compare the LG length centimorgan distance between LGs (chromosome size), number of crossovers (COs), and SNP number and bin per LG. In addition, correlations were performed using PROC CORR of SAS, Pearson correlation (ver. 9.4; SAS Institute, Cary, NC, USA).

After analyses and LG development, previously determined LGs[21] in the work by Hulse-Kemp et al were used to validate and to determine orientation. These LGs were also aligned based on the cotton genome sequence reference. Given the preserved subgenome separation of the ancestral origin (A versus D) of specific chromosomes, researchers have numbered them accordingly, denoting the 13 A-derived chromosomes as 1 to 13 (the A_t subgenome) and the 13 D-derived

chromosomes as 14 to 26 (the D_t subgenome). A colinear one-to-one relationship between all A- and D-subgenome chromosomes is not possible, even considering only the major structural differences (historical translocations). As a working model, we have used the conventional nomenclature for chromosomes (c01-c26) because it avoids inadvertently indicating chromosome-wide homeology where it is lacking. For the purposes of this report, we will designate chromosomes and the corresponding LGs similarly, eg, c01 and c02 for chromosomes (LGs) 1 and 2. When necessary, we state chromosome (Chr) or LG. In addition, through the text, we use SNP bins or genetic bins which are defined as more than 1 SNP or a series of SNPs with identical genotypes and thus fell at the same position in the genetic map(s).

Comparative genomics and syntenic analyses

All 50-bp (base pair) probe sequences used to assay SNPs on the array were aligned to the NBI *G hirsutum* L. acc. TM-1 $(AD)_1$[39] and to the high-quality JGI *G raimondii* (D_5) reference genome[36] using Bowtie2, requiring a full 50-bp alignment to a unique position in the genome without any mismatches.[46] Map positions were plotted against AD_1 and D_5 alignment positions for mapped markers. Positions from the F_2 map in this study were also plotted against previously reported positions in the F_2 map. In addition, all SNP marker sequences from the consensus map were aligned to the NBI *G hirsutum* L. acc. TM-1 genome reference using CLC Genomics Workbench 8.5.1 (www.clcbio.com) basic local alignment search tool (BLAST) with 50-word size, at least 95% DNA sequence similarity and $E < 1.0 \times E{-}5$.

Results

SNP maker segregation

The Illumina Infinium genotyping CottonSNP63K array and cluster files provided a total of 7417 genotyped SNP markers on 3 full-sib intraspecific (*G hirsutum* × *G hirsutum*) independent mapping populations derived from the cultivars "Phytogen 72" (PHY72) and "Stoneville 474" (STV474), which included 93 F_2s from PHY72 × STV474, 132 RILs from PHY72 × STV474 and 104 reciprocal RILs from STV474 × PHY72. After filtering, a total of 7171 SNPs were obtained in the F_2 population; 7172 SNPs in the RIL population; and 6605 SNPs in the reciprocal RIL population with only 4.0%, 3.2%, and 3.3% of segregation distortion observed in these SNP data sets, respectively ($P > .05$). When we examined the distortion for each LG or chromosome (c) within and across populations, c17 (20.1%), c22 (16.1%), and c25 (12.8%) from the D_t subgenome contained some highly distorted SNPs only in the F_2 population, whereas c04 (ranked from 5.2% to 12.6%), c6 (ranked from 13.1% to 24.7%), c07 (ranked from 6.3% to 12.6%), and c09 (ranked from 2.9% to 3.7%) contained some distorted SNPs in all 3 populations. As expected, most of the markers incorporated into the genetic linkage maps from all populations were primarily derived from the intraspecific *G hirsutum*–designated content on the CottonSNP63K array rather than the interspecific content from other species: *G barbadense, Gossypium tomentosum, Gossypium mustelinum, Gossypium longicalyx*, or *Gossypium armourianum* (Supplementary Table S1).

Genetic linkage maps

To reduce linkage mapping errors and minor SNP order differences on a LG or chromosome during the analysis, we used a similar approach to develop the group nodes (*independence LOD*) and maps (maximum likelihood) and cutoffs/thresholds for grouping and linkage between 2 SNP markers (Kosambi with maximum distance of 40 cM) in all populations. Most of the SNP markers used to construct the linkage maps showed normal mendelian segregation and were assimilated into 26 LGs corresponding to the 26 cotton chromosomes (n = 26). Three genetic linkage maps were constructed from the 3 full-sib intraspecific populations using the maximum likelihood algorithm (Table 1).

In the F_2 PHY72 × STV474 population, 7034 SNPs were grouped on 26 different chromosomes. The high-density map comprised 2212 genetic bins and covered 3597 cM distance of the cotton genome with an average SNP interval between 2 linked markers of 1.8 cM (Table 1). The largest per chromosome average distance between 2 SNPs or interval gap was observed on c11 (3.25 cM), which also was the second largest LG (211.06 cM). The A_t-subgenome LGs contained fewer bins (1033) than the D_t subgenome (1179), which reflected the higher number of SNP markers (9.1%) in the D_t subgenome and slightly high rates of recombination (Avg. = 2.61 COs). The highest segregation distortion was observed in c17, followed by c22 and c25 (Supplementary Table S2). The F_2 population averaged 49.4% heterozygous loci with the c12 and c26 homeologous pair having the highest percentages of heterozygous loci at 52.5% and 54.2%, followed by c02 (53.9%) and c06 (53.5%) (Figure 1). The LGs of this F_2 population averaged 138 cM. The longest members the A_t and D_t subgenomes, respectively, were homeologs c05 (225 cM) and c19 (203 cM), which had high average CO events 4.19 and 2.92, respectively (Figure 2).

In the RIL PHY72 × STV474 population, 7059 SNPs were grouped on 26 different chromosomes. The high-density map comprised 2620 genetic bins and covered 3966 cM of the cotton genome with an average SNP interval between 2 linked markers of 1.6 cM (Table 1). Similar to the F_2 population, the largest per chromosome average distance between 2 SNPs occurred on c11 (2.8 cM). Fewer genetic bins also occurred in the A_t subgenome (1211) than in the D_t subgenome (1409). The D_t subgenome also had 8% more SNPs with the rate average slightly high rates of CO (5.05). And similar to the F_2, these high number of SNPs in the D_t subgenome may have

Table 1. Distribution of 7322 single-nucleotide polymorphism (SNP) marker loci distributed across the 26 allotetraploid cotton (*Gossypium hirsutum* L.) chromosomes on 3 populations (1 F_2 and 2 recombinant inbred line [RIL]) derived from parental cultivars Phytogen 72 (Phy72) and Stoneville 474 (STV474).

CHROMOSOME (C)	93 F_2 PHY72×STV474				132 RIL PHY72×STV474				104 RIL STV474×PHY72			
	NO. OF SNPS	SNP BIN	SIZE, CM	AVG. SNP INTER. CM	NO. OF SNPS	SNP BIN	SIZE, CM	AVG. SNP INTER. CM	NO. OF SNPS	SNP BIN	SIZE, CM	AVG. SNP INTER. CM
A_t subgenome												
c01 (A01)	194	64	129.93	2.03	192	89	142.32	1.60	176	68	148.92	2.19
c02 (A02)	213	67	111.12	1.66	217	92	134.14	1.46	195	79	138.76	1.76
c03 (A03)	226	79	136.91	1.73	225	97	132.16	1.36	193	85	101.14	1.19
c04 (A04)	134	55	87.62	1.59	128	71	120.10	1.69	111	64	121.24	1.89
c05 (A05)	348	142	224.59	1.58	364	136	235.79	1.73	277	140	239.27	1.71
c06 (A06)	73	30	72.98	2.43	89	46	100.95	2.19	84	50	108.56	2.17
c07 (A07)	230	80	148.63	1.86	227	87	171.33	1.97	190	72	109.73	1.52
c08 (A08)	363	108	160.12	1.48	360	140	142.37	1.02	315	118	165.19	1.40
c09 (A09)	140	60	165.14	2.75	142	73	186.45	2.55	135	67	191.15	2.85
c10 (A10)	182	60	77.97	1.30	217	90	157.76	1.75	166	65	101.82	1.57
c11 (A11)	148	65	211.06	3.25	131	61	170.42	2.79	120	63	199.53	3.17
c12 (A12)	228	61	129.88	2.13	233	80	121.69	1.52	216	85	136.57	1.61
c13 (A13)	720	162	131.89	0.81	723	149	152.86	1.03	629	192	200.16	1.04
Subtotal A_t	3199	1033	1787.85	1.89	3248	1211	1968.34	1.74	2807	1148	1962.02	1.85

(Continued)

Table 1. (Continued)

CHROMOSOME (C)	93 F_2 PHY72×STV474				132 RIL PHY72×STV474				104 RIL STV474×PHY72			
	NO. OF SNPS	SNP BIN	SIZE, CM	AVG. SNP INTER. CM	NO. OF SNPS	SNP BIN	SIZE, CM	AVG. SNP INTER. CM	NO. OF SNPS	SNP BIN	SIZE, CM	AVG. SNP INTER. CM
D_t subgenome												
c15 (D01)	272	63	157.13	2.49	266	84	142.45	1.70	242	66	150.27	2.28
c14 (D02)	451	116	130.65	1.17	452	132	135.88	1.03	418	128	138.03	1.08
c17 (D03)	229	79	124.56	1.58	233	119	120.59	1.01	213	87	131.17	1.51
c22 (D04)	205	77	120.54	1.57	202	89	129.93	1.46	182	79	115.62	1.46
c19 (D05)	329	157	203.23	1.29	337	143	221.45	1.55	326	138	200.31	1.45
c25 (D06)	273	81	113.08	1.40	271	107	149.15	1.39	253	100	98.07	0.98
c16 (D07)	506	128	148.22	1.16	508	145	147.50	1.02	470	146	165.65	1.13
c24 (D08)	576	132	163.30	1.24	558	167	170.89	1.02	501	152	169.50	1.12
c23 (D09)	223	78	128.06	1.64	220	91	146.11	1.61	198	87	156.52	1.80
c20 (D10)	267	90	116.21	1.29	264	110	144.28	1.31	241	93	137.85	1.48
c21 (D11)	159	47	124.34	2.65	164	67	171.75	2.56	151	50	151.14	3.02
c26 (D12)	152	65	154.30	2.37	151	74	163.19	2.21	144	73	157.66	2.16
c18 (D13)	193	66	126.06	1.91	185	81	154.62	1.91	174	71	127.92	1.80
Subtotal D_t	3835	1179	1809.68	1.67	3811	1409	1997.79	1.52	3513	1270	1899.71	1.64
Total	7034	2212	3597.53	1.78	7059	2620	3966.13	1.63	6320	2418	3861.74	1.74

Avg. SNP inter. cM, average cM interval distance between 2 linked SNP markers.
The percentage of similar or common SNP markers between each of the populations ranged from 96% to 99%, with the highest number of different or unique SNPs observed on the reciprocal RIL population STV474×PHY72. c13, c14, and c24 had most of the unique mapped in a specific population.

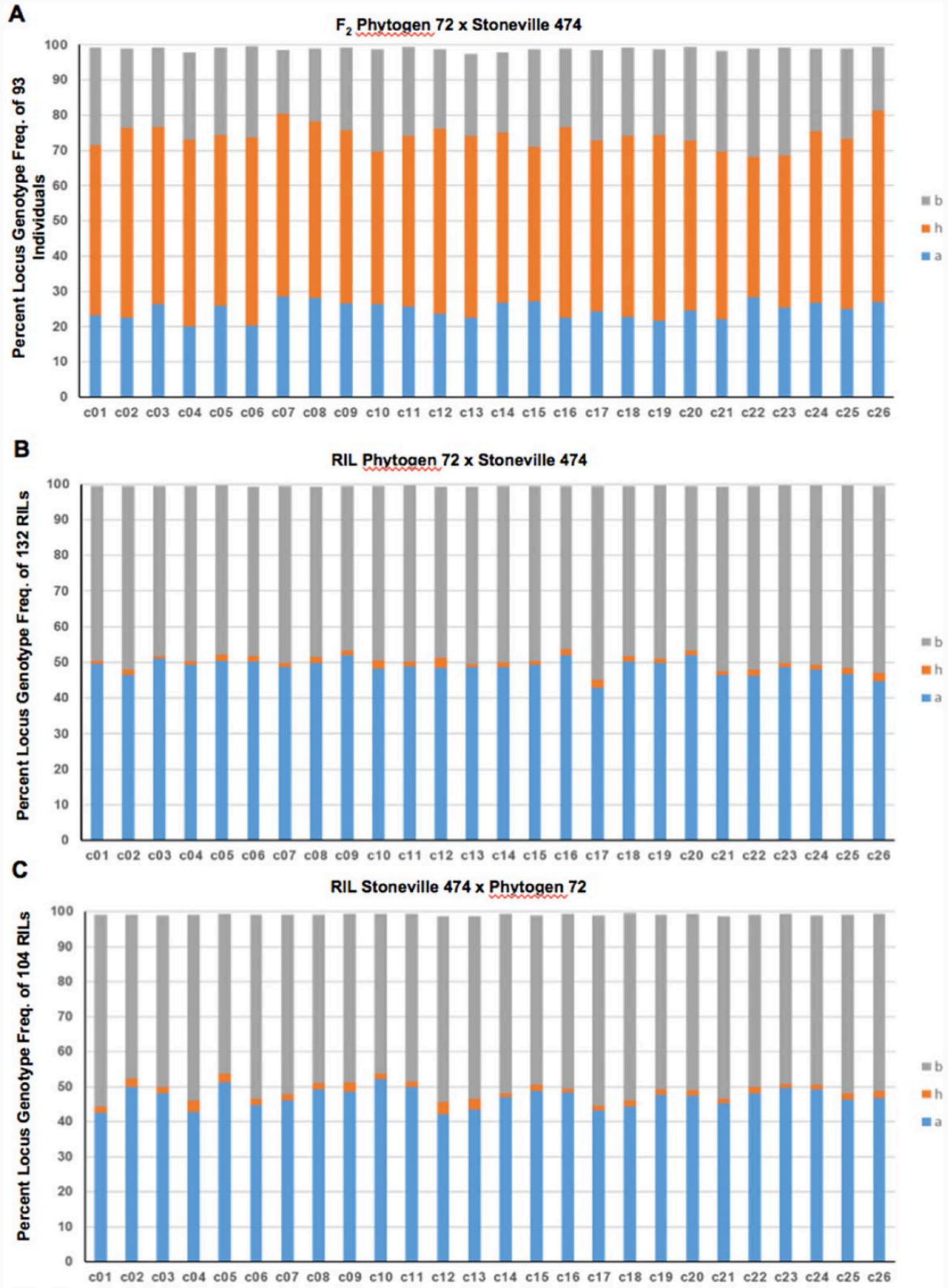

Figure 1. Distribution of genotypes in assessed populations, F_2 Phytogen 72 (PHY72) × Stoneville 474 (STV474) with 93 individuals, RIL PHY72 × STV474 with 132 lines, and RIL STV474 × PHY72 with 104 lines, exhibiting heterozygote loci. RIL indicates recombinant inbred line. Chromosome (c), a = PHY72 SNP allele, b = STV474 SNP allele, and h = heterozygous.

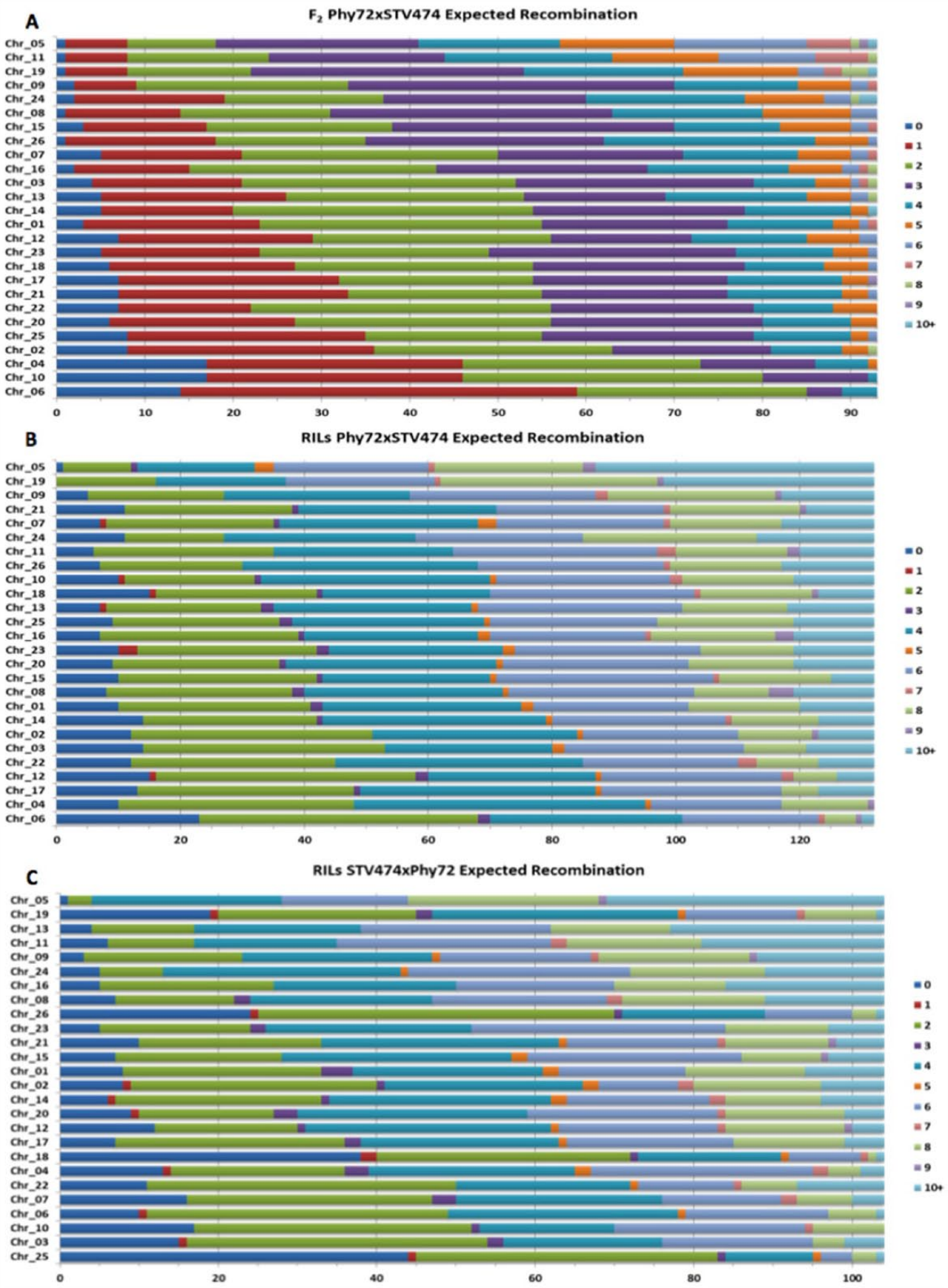

Figure 2. Distribution of expected recombination frequency or crossovers per chromosome from 3 populations, F_2 Phytogen 72 (PHY72) × Stoneville 474 (STV474) with 93 individuals, RIL PHY72 × STV474 with 132 lines, and RIL STV474 × PHY72 with 104 lines. Chromosomes are arranged from small to large based on centimorgan length distance. RIL indicates recombinant inbred line. Y axis = chromosome (Chr), X axis = number lines, and 0 to 10+ = average number of recombination events or cross overs.

Table 2. Distribution of single-nucleotide polymorphism (SNP) marker loci across the 26 allotetraploid cotton (*Gossypium hirsutum* L.) chromosomes on the consensus or JoinMap derived from 3 populations from parental cultivars Phytogen 72 (Phy72) and Stoneville 474 (STV474) (93 F_2 Phy72 × STV474, 132 RIL Phy72 × STV474, and 104 RIL STV474 × Phy72 populations).

CHROMOSOME (C)	NO. OF SNPS	SNP BIN	SIZE, CM	AVG. SNP INTER. CM	GAPS >10 CM
A_t subgenome					
c01 (A01)	199	130	131.78	1.01	1
c02 (A02)	223	122	122.63	1.01	2
c03 (A03)	229	150	143.84	0.96	1
c04 (A04)	134	97	118.29	1.22	2
c05 (A05)	368	229	233.67	1.02	0
c06 (A06)	91	72	88.26	1.23	1
c07 (A07)	237	141	123.95	0.88	1
c08 (A08)	373	207	129.36	0.62	0
c09 (A09)	144	107	154.81	1.45	2
c10 (A10)	225	134	126.01	0.94	2
c11 (A11)	163	107	230.87	2.16	5
c12 (A12)	239	109	97.43	0.89	1
c13 (A13)	731	178	123.78	0.70	0
Subtotal A_t	3356	1783	1824.68	1.08	1.4
D_t subgenome					
c15 (D01)	276	130	126.14	0.97	1
c14 (D02)	447	177	115.60	0.65	2
c17 (D03)	238	160	109.51	0.68	0
c22 (D04)	206	131	107.35	0.82	0
c19 (D05)	344	236	209.80	0.89	0
c25 (D06)	278	161	106.88	0.66	1
c16 (D07)	523	243	133.26	0.55	0
c24 (D08)	572	200	139.19	0.70	1
c23 (D09)	225	145	127.90	0.88	1
c20 (D10)	271	160	124.47	0.78	1
c21 (D11)	166	80	144.06	1.80	3
c26 (D12)	155	114	145.30	1.27	1
c18 (D13)	187	104	123.64	1.19	2
Subtotal D_t	3888	2041	1713.10	0.91	1
Total	7244	3824	3537.78	1.00	1.2

influenced bin detection and bin number. The highest segregation distortion occurred on c06, followed by c26 (Supplementary Table S3). The RIL population averaged 1.5% heterozygous loci, with the highest levels occurring in homeologous chromosomes c12 and c26 (3.0% and 2.4%), followed by c10 (3.0%) and c25 (3.2%) (Figure 1). The LGs of this RIL population averaged 153 cM, and subgenome averages were nearly identical, too; this seems to support the hypothesis that subgenomic differences in bin number arose due to subgenomic differences in numbers of SNPs. The longest members in the RIL of the

A_t and D_t subgenomes, respectively, were also homeologs c05 (236 cM) and c19 (221 cM), which also had the highest average CO events 7.93 and 7.46, respectively (Figure 2).

In the reciprocal RIL STV474 × PHY72 population, 6320 SNPs were grouped on 26 different chromosomes. The high-density map comprised 2418 SNP bins and covered 3862 cM of the cotton genome with an average SNP interval between 2 linked markers of 1.7 cM. Similar to the F_2 population, the largest average SNP interval gap was observed in c11 (3.2 cM). As for the other 2 populations, fewer bins occurred in the A_t subgenome (1148) than the D_t subgenome (1270), possibly because there were 11.2% more SNP markers in the D_t subgenome. However, the A_t subgenome with less SNPs averaged slightly high rates of CO (4.97). The highest segregation distortion occurred in c07 followed by c01 (Supplementary Table S4). This reciprocal RIL population averaged 1.9% heterozygous loci, and the highest rates of heterozygous loci were observed for homeologous chromosomes c12 and c26 (3.7% and 3.1%, respectively, Figure 1). The LGs of this RIL population averaged 149 cM. As for the 2 other populations, the longest LGs in the A_t and D_t subgenomes, respectively, were homeologs c05 (239 cM) and c19 (200 cM), which also had high average CO events 8.20 and 3.46, respectively. The next longest were nonhomeologs c13 (200 cM) and c24 (172 cM). Both RIL populations showed slightly higher overall rates of recombination (3966 and 3862 cM) than the F_2 population (3598 cM, Figure 2).

When we further examined rates of recombination or CO, chromosome size, and SNP and genetic bin number per chromosome and between subgenomes, significant differences were observed for individual LG map lengths or chromosome size for the 3 populations with a mean LSD of 28.0 cM between 2 LGs. Significant differences were also observed for SNP number and bin number and CO between 2 LGs ($P > .05$). Even though in the 3 populations the A_t-subgenome LGs contained fewer SNPs and genetic bins than the D_t subgenome, no significant differences were observed between these 2 subgenomes for the above events. The subgenome average distances of cM were nearly identical. The correlation between SNPs and detection of bins was $r = .87$. This correlation further supports the hypothesis that subgenomic differences in bin number arose due to subgenomic differences in numbers of SNPs. In addition, in this study, the correlation for average CO and LG length from the 3 populations was $r = .70$, and for average CO and SNP bin was $r = .51$ ($P > .05$). However, examination of CO of the 2 subgenomes within each population revealed that SNP number or genetic bins did not affect CO. However, in the populations with the same female-cross F_2 PHY72 × STV474 (2.60 COs) and RIL PHY72 × STV474 (5.06 COs), the D_t subgenome revealed slightly high average rates of recombination, whereas in the reciprocal RIL STV474 × PHY72 (4.97 COs), the A_t subgenome revealed slightly high average rates of recombination per chromosome, indicating that CO did not depend on SNP number or genetic bins.

A consensus map was assembled from the 3 independent population genetic linkage maps with a total of 329 progeny (F_2s and RILs) and assimilated a total of 7244 SNP markers (Table 2). For map or group integration, only the regression mapping algorithm is available in the JoinMap program. The high-density consensus map comprised 3824 genetic bins (7244 SNP markers) and covered 3538 cM of the cotton genome with an average SNP interval between 2 linked markers of 1.0 cM (Supplementary Table S5 and Table 2). As was found in all individual populations, the largest average distance between 2 SNPs was observed in c11 (2.2 cM), and the fewer bins (46.6%) occurred in the A_t subgenome (1783) than in the D_t subgenome (2041). The D_t-subgenome consensus map LGs included 7.3% more SNP markers than the A_t subgenome, whereas the overall lengths of the respective LGs, the A_t and D_t subgenomes accounted for similar percentages of the estimated recombination, 51.6% and 48.4%, respectively. The shortest LGs were the homeologous chromosomes c06 and c25, whereas the longest were the segmentally homeologous pair c05 and c19 (Table 2). The consensus map LGs averaged 136 cM, and subgenome averages (140 and 132 cM) were nearly identical. As for the 3 other populations, the longest LGs in the A_t and D_t subgenomes, respectively, were homeologs c05 (234 cM) and c19 (210 cM), followed by c11 (231) and c21 (144 cM, Figure 2). The shortest LGs were homeologs c06 (88 cM) and c25 (107 cM). The overall consensus map (3538 cM) was close to the overall length of the F_2-based map (3598 cM). On average, 4.0 COs were exhibited on the 26 chromosomes of the upland genome with homeologs c05 and c19 exhibiting the highest average number of CO (6.8 and 4.9), followed by c11 (Figure S1) with a 5.1 average (Figure 3).

Even though the number of identical SNP markers varied based on recombination frequencies, for the most part, the SNPs were generally grouped or derived from the same LG or chromosome in each population. However, there were SNPs that were only genotyped/mapped in one of the populations possibly because of filtering of individual SNPs or different CO events in each population. The percentage of similar or common SNPs between each of the populations ranged from 96% to 99%, with the highest number of different or unique SNPs observed on the reciprocal RIL population STV474 × PHY72. Overall, the SNP makers of the new consensus map aligned well with the previous published F_2 map (Figure S2) and with all of the developed maps. Grouping, linkage, and gene-SNP marker order showed consistency across LGs or segmental homology for the 3 mapping populations as represented for c04 and c22 (Figures S3A and S3B). The consensus map contains 99% of all the mapped SNP markers from all 3 populations assimilated into 26 LGs. Linkage groups c13, c14, and c24 of the consensus map did not include 24, 18, 10 SNP markers, respectively, that had been uniquely mapped in a specific population. By capturing the CO events of the 329 progeny from these populations, the

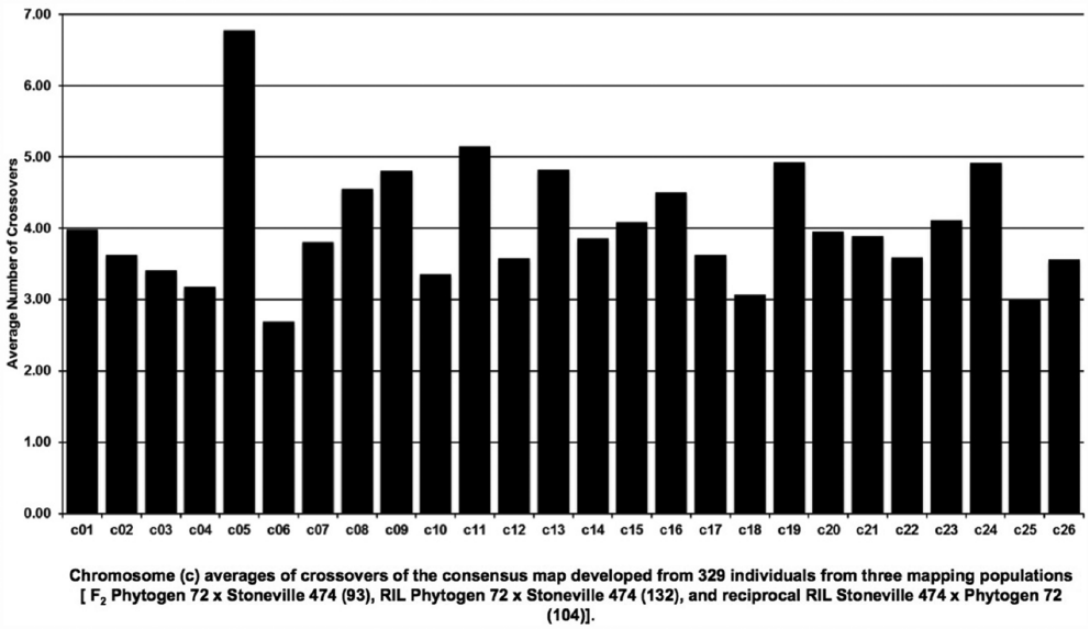

Figure 3. Distribution of expected recombination frequency or crossover average of chromosomes from the consensus map developed with 329 individuals from 3 mapping populations (F_2 Phytogen 72 × Stoneville 474 [93 individuals], RIL Phytogen 72 × Stoneville 474 [132 individuals], and RIL Stoneville 474 × Phytogen 72 [104 individuals]). Linkage groups (A and D)/chromosomes (c). RIL indicates recombinant inbred line.

consensus map increased the number of bins and SNP markers mapped to the 2 subgenomes, provided much better placement of gene/SNP marker order, and improved the coverage or distribution of SNPs through the *G hirsutum* genome (Tables 2 and Supplementary Table S5, Figure S3).

Comparative genomic and syntenic analyses

All of the 3824 bins and corresponding DNA-derived SNP sequence markers on the consensus map were aligned to the NBI *G hirsutum* acc. TM-1 AD_1 reference genome[39] using Bowtie2. Synteny was detected for the total 3824 mapped SNPs (Figure 4). The genetic linkage consensus map versus the NBI *G hirsutum* acc. TM-1 AD_1 reference genome (Figure 4) and the D_5 reference genome also showed high collinearity across LGs and chromosomes, with a higher resolution and with increased number of genetic bins. In addition, linkage map positions for the SNP markers on the previously published F_2 PHY72 × STV474 map[21] were compared with the marker positions of the consensus map. The F_2 map previously published versus the linkage consensus map showed high collinearity across the 26 LGs (Figure S2).

The genetic linkage consensus map versus the NBI *G hirsutum* acc. TM-1 AD_1 reference genome (Figure 4) and the D_5 reference genome also showed high collinearity across LGs and chromosomes, with a higher resolution and with increased number of bins. Sequence alignment or blast analyses using CLC genomics of the 3824 bin DNA–derived SNP sequence markers of the consensus map revealed sequence homology to the NBI *G hirsutum* acc. TM-1 AD_1 reference genome (Supplementary Table S6). A moderate number of sequences of SNPs that were linkage mapped to the A_t-subgenome LGs were found by sequence alignment to associate with a NBI D_t-subgenome scaffold but not the A_t subgenome with both sequence alignment methods (Figure 4, Supplementary Table S6). All A_t-subgenome (c1-c13) chromosomes showed a variable number of associated marker sequences having homology to the corresponding homeologous chromosome from the D_t subgenome (c14-c26). The same phenomenon appeared with SNPs that linkage mapped to the Dt subgenome, however, on a much lower level. The percentages of SNP sequences that aligned to the homeolog in the opposite subgenome to which they were linkage mapped ranged from 24% (c7) to 48% (c11) in the A_t subgenome. In addition, SNP sequences from 4 chromosomes c2, c03, c04, and c05 had hits to more than one homeolog-pair of the D_t-subgenome reference (Supplementary Table S6), confirming the historical translocations occurring among c2/c03 and c04/c05. The BLAST alignments were also able to detect 364 SNP-associated to unintegrated scaffolds. These 364 SNP-associated unintegrated scaffolds can be placed onto pseudochromosomes of A_t and D_t subgenomes of the NBI *G hirsutum* assembly, prospectively increasing coverage by the 47.7 Mb. Moreover, 112 and the 364 unintegrated scaffolds were for the first time identified with SNP markers in this study to belong to the A_t subgenome and 89 scaffolds to belong to the D_t-subgenome reference and may be placed on specific cotton chromosomes, increasing *G hirsutum* genome reference coverage by 2.4% (Supplementary Table S6).

Discussion

The independent high-density intraspecific genetic linkage maps and consensus map developed with the CottonSNP63K

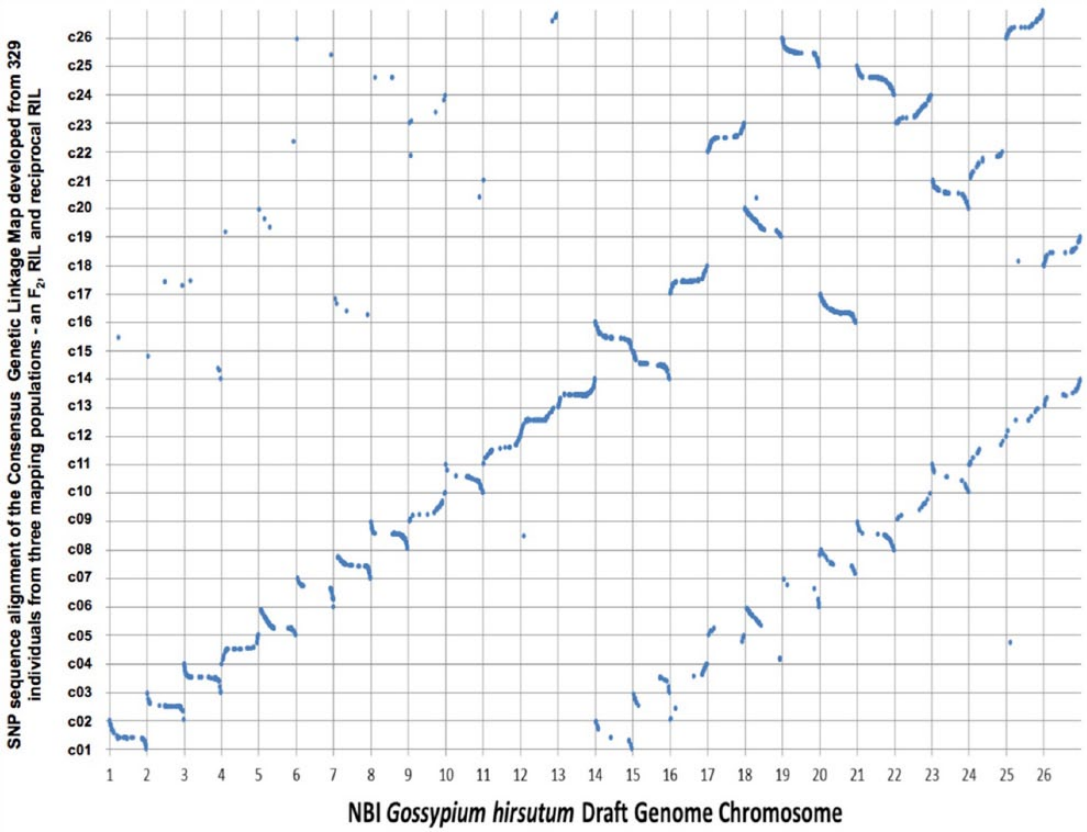

Figure 4. Dot plot of the syntenic position of SNP markers in the allotetraploid interspecific genetic linkage consensus map versus the NBI *Gossypium hirsutum* L. reference genome. The 26 allotetraploid chromosomes are shown on the *x*-axis and the 26 linkage groups of the consensus map are shown on the *y*-axis showing 3824 mapped markers. SNP indicates single-nucleotide polymorphism.

array represent a valuable resource which will help to advance genetic improvements needed in the allotetraploid $(AD)_1$ upland cotton (*G hirsutum*). The CottonSNP63K array enabled expedient development of high-quality, high-density maps with more than 7000 scorable polymorphic SNPs between PHY72 and STV474 cultivars and advanced our understanding of upland cotton genetic recombination by examining parental relationships, segregation and gene/SNP marker order from F_2 population to F_7 generation, and genome organization of the cotton crop. The 3 different populations (a F_2, a RIL, and a reciprocal RIL population) provide a robust biparental platform for follow-up research for genetic analysis. By examining placement of SNP and bin number, chromosome size, and rates of recombination or CO on the 26 chromosomes and between subgenomes (A_t 1-13 and D_t 14-26), we increased our insight of paleopolyploidy-derived genomic complexities of this valuable natural fiber and oil crop.

In all 3 populations, the SNPs were assimilated into 26 LGs that correspond to the 26 chromosomes of the cotton genome. And the intraspecific consensus map is the first assembled in upland cotton using a core of SNP markers assayed on different cotton populations derived from 2 cultivars with distinctly different genetic backgrounds, yield, and fiber quality. This map increased the number of bins and SNP markers mapped to the 2 subgenomes, provided much better placement of gene/SNP marker order, and improved the coverage or distribution of SNPs through the *G hirsutum* genome. The high-density genetic consensus map of the upland allotetraploid comprised 3824 genetic bins with similar recombination frequencies in the 2 subgenomes (A_t and D_t). The estimated genome coverage of all maps in this study ranged from 3537 cM (regression mapping algorithm—linkage JoinMap) to 3966 cM (maximum likelihood mapping algorithm RIL PHY72 × STV474 map) (Tables 2 and 3). They fall well within the range of previous map sizes of published interspecific maps 3380 to 5115 cM[20,22,42,47] and within 2061 to 4448 cM of reported intraspecific maps.[5,18,27,48] In addition, the developed maps in this study, together with the recently published *G hirsutum* L. acc. TM-1 genome references[38,39] and ancestor diploid (JGI *G raimondii*[36] and BGI *G arboreum*[35]), reference genomes provide a foundation for fine mapping and genetic dissection of candidate genes and QTL for agronomically important traits such as yield and fiber quality traits, drought and plant stress tolerance, and pest and disease resistance. In addition, it will also foster map-based cloning and genome assembly efforts, as well as contribute to advancements in marker-assisted selection and genomic selection in upland breeding programs.

Polyploidy is a common event now recognized in all angiosperm genomes. During the process of speciation and then after, the allotetraploid crop such as cotton experienced

genome merging, DNA duplication, and non-mendelian interaction and processes. Then, alteration of activation of genes, retroelements, and several kinds of homeologous interactions and exchanges occur thereafter.[49] The extant of the ancestral A genomes (A_1 and/or A_2) is about twice the size and of much greater complexity than the extant D genomes. By capturing the CO events of the 329 progeny (F_2 and RILs) from these populations, we were able to further examine placement/map on chromosomes SNPs/loci and genetic bin number, chromosome sizes, and rates of recombination or CO per LG and between subgenomes (A_t and D_t). The patterns of chromosome-specific variations were largely consistent across mapping populations. From the ANOVA, significant differences were observed for individual LG map lengths for the 3 populations with a mean LSD of 28.0 cM between 2 LGs. In addition, significant differences were observed for SNP number and bins and average CO between 2 LGs ($P > .05$). The present maps and consensus map revealed that more SNP loci were mapped (ranged from 91 to 731) to D_t subgenome, resulting in a slightly shorter average distance between 2 markers and a low number of gaps (>10) in this subgenome, consistent with previously published research.[19,21,22] The correlation between LG lengths of specific chromosomes in the consensus map versus the 3 LGs of the populations follows the pattern of chromosome-specific differences of LG lengths with $r = .85$, $r = .82$, and $r = .76$, respectively (Figure 2). Even though the overall LG average length distances of these mapping populations differed ($F_2 = 138$ cM and reciprocal-RIL = 152 cM), the subgenome average distances were nearly identical, indicating that subgenomic differences in bin number arose due to numbers of SNPs.

The overall recombination rate in cotton have been reported in the range of 0.5 cM per 1 Mb to 5.7 cM per 1 Mb with an average of 1.75 cM per 1 Mb.[22] In this study, a preliminary examination of the most even lengthwise homeologs (c01/130 cM, A_t subgenome and c15/130 cM, D_t subgenome) of the consensus map with 23 SNP markers (11 SNPs/c01 and 12 SNPs/c15) spaced at 1 cM at different spots through the LGs revealed recombination averaging around 1 cM per 0.5 Mb for the A_t subgenome and 1 cM per 0.2 Mb for the D_t subgenome in this consensus LGs. In addition, even though in the 3 populations the A_t-subgenome LGs contained fewer SNPs and genetic bins than the D_t subgenome, no significant differences were observed between these 2 subgenomes for the above expected CO. However, based on SNP loci and CO, paleopolyploidy-derived genomic complexities were observed. Recent studies[22,49,50] based on DNA sequences and RNAseq transcriptome analyses have reported bias on gene duplication and expression levels in allotetraploid crops, including cotton. Herein, homeolog bias is referred to the preference for high number of expected recombination events or CO of LG(s)/chromosome(s) and subgenomes. c11 and c05 from the A_t subgenome and c24 from the D_t subgenome exhibited high to slightly high CO on the 3 populations. The number of CO per chromosome is moderately correlated ($R^2 = 0.70$-0.79) with genetic LG size.[21] The larger homeologs, c05 (6.8) and c19 (4.9), had a high number of CO, followed by c11 with 5.1 average of CO (Figure 3). However, examination of CO of the 2 subgenomes within each population revealed that COs were not affected by the SNPs or SNP bins in these subgenomes.

Another interesting phenomenon observed in this study was preferential expected recombination events between subgenomes. In gene expression analyses of allotetraploid hybrids, similar phenomenon is described as "parental dominance," in which slight overall expression levels favored one of the parents.[50,51] In this study, the D_t subgenome in the same female-cross F_2 PHY72 × STV474 (2.6 COs) and RIL PHY72 × STV474 (5.1 COs) exhibited slightly high average rates of recombination compared with the A_t subgenome 2.5 and 4.8, respectively, whereas in the reciprocal RIL STV474 × PHY72, the A_t subgenome exhibited slightly high average rates (4.9 COs) compared with the Dt subgenome (4.1 COs). Overall recombination was higher (10.2% and 7.3%) in the RIL populations (3966 and 3862 cM) than in the F_2 (3598). Most or all of the recombinant phenotypes produced in a biparental cross were captured in suitably sized F_2 generation, and these populations are efficient for mapping and examining high-heritable traits.

Recombinant inbred line populations are most suitable for complex traits in which replicated tests, multiyear, and multilocation experiments are needed. In this study, the F_2 allotetraploid population revealed a high number of recombinant individual genotypes. And through the successive generations of the RIL populations, we were able to capture and maintain a high number of recombinant genotypes. These recombinant RILs were confirmed with this SNP marker set. The F_2 and the 2 RIL populations provide a robust valuable biparental resource for upland cotton. In addition, with the genome SNP coverage and rates of recombination of the 26 cotton chromosomes, this study provides additional insight in understanding trait inheritance during the breeding process.

Knowing that our data sets in all 3 populations contained some distorted SNPs in some LGs or chromosomes (2.9%-24.7%) and heterozygote loci in the RIL populations (PHY72 × STV474 [1.5%] and STV474 × PHY72 [1.9%]), we further examined the LGs for CO interferences and/or SNP-calling errors which can result in change of marker order and increased COs and map sizes. In each of the populations, a few of the LGs showed a few SNP markers with CO interferences (data not shown). By manually removing a SNP marker with more than 20 SNP interference positions or replacing the data point as missing data point for each interference in a LG, in subsequent analyses, we observed that SNP marker order did not change. However, LG sizes may be varied from around 3 to 15 cM of total distance of the examined LGs. The reciprocal RIL STV474 × PHY72 with 1.9% heterozygote loci had

the highest number of LGs (c2, c5, c7, c12, c14, c16, c22, and c24) with interferences. However, this RIL population produced the lowest total map size distance (3861.74 versus 3.966.13 cM) of the 2 RIL populations (Table 1). As we all know, no mapping program can ever produce the ultimate genetic map, and the selection of subsets of loci and individuals of any giving project will dictate quality of the produced linkage maps (JoinMap user manual). Our Illumina data were of high quality and our mapping approach also was able to produce small LG cM distance total size for c06 (72.98 cM) and c12 (129.88 cM) compared with the previously intraspecific F_2 PHY72 × STV474 LGs (c06 = 111.0 cM and c12 = 179.0 cM).[21] Additional research is needed to resolve CO interferences and/or SNP-calling errors in these large mapping projects. In our laboratories, research is ongoing and at least 2 additional intraspecific and 3 interspecific mapping populations are being genotyped with the Cotton63k array to provide additional resources and an ever stronger platform for localizing and identifying agronomically important loci for the improvement of the cotton crop.

Mapping by meiotic configuration analysis placed the ancestral c02 and c03 break points extremely close (circa 1 cM) to the respective centromeres and those of c04 and c05 near their respective centromeres (circa 6 or 7 cM).[52] It has been suggested that these ancient intra-A-subgenomic translocations involved complete arms.[42] The break points for the translocations were roughly mapped using genetic linkage and physical maps to the D_5 chromosomes.[21,22] Even though not fully addressed in this study, comparative genomic analyses revealed the 2 previously reported intra-A_t-subgenomic reciprocal translocation events that affect the structure of extant cotton chromosomes c02 and c03, as well as chromosomes c04 and c05, relative to the D_t subgenome and the genomes of extant diploid A and D genome species. These affect homeologous relationships with D_t-subgenome chromosomes c14 and c17, as well as c19 and c22, too. In addition to these historically recognized ancestral A-subgenome reciprocal translocations, 15 simple translocations on these subgenomes have also been reported along with 19 possible inversions[22] which are slightly different from previously reported research.[5,37] Similar chromosome rearrangements were observed in the syntenic dot plots with insertions of mapped SNP marker sequences of the A_t subgenome observed and sequences relocated on chromosomes of the *D_{t-1}-Subgenome* (Figure 4). Assuming that the NBI *G hirsutum* genome reference and its orientation is mostly correct, our comparative genomic analyses also revealed evidence of additional unconfirmed possible duplications, inversions and translocations, and unbalance SNP sequence homology (ranging from 24% to 48%) or SNP sequence/loci genomic dominance, or homeolog loci bias of the upland tetraploid A_t and D_t subgenomes[49] (Supplementary Table S6 and Figure 4). Given the challenges of short-read NGS assemblies, it might be prudent to approach such interpretations with caution until strongly confirming data can be obtained.

Based on the alignment of SNP sequences in the 3824 genetic bins of the genetic linkage consensus map, syntenic analyses revealed high collinearity with available related genome sequences such as the NBI *G hirsutum* L. acc. TM-1 (Figure 4) and the JGI D_5[36] genome references.[21] These genomic analyses also provided some additional insight into structural variation and localization information in the allotetraploid upland cotton genome. From the sequence alignment analyses, a total of 364 SNP-associated unintegrated scaffolds were identified, increasing the coverage of the upland genome reference overall of total sequence size by 2.44% (Supplementary Table S6).

This first high-density SNP genetic linkage consensus map represents a valuable resource for *G hirsutum*. With a core of reproducible mendelian SNP markers assayed on different intraspecific populations from crosses involving the same parents, these population-specific maps and the consensus map are resources for subsequent genetic research, genome analysis, and breeding. They will facilitate future genome assemblies, including the integration of sequenced physical mapping resources. Given the tremendous utility of RIL populations for analysis of multiple complex trait analysis across diverse replicated experiment locations, the maps and SNP-genotyped RILs will be extremely useful for identifying QTLs defined by genetic differences between these 2 parents. These are likely to include traits for agronomic and physiological improvements, abiotic stress, disease and pest resistances, and enhanced fiber attributes. This research provided further knowledge of parental relationships, gene order, and insights into genetic recombination and genome organization of the cotton crop.

Acknowledgements

The authors would like to thank D Laumbach, J Sanchez and students working at the USDA-ARS, Lubbock, TX, for assisting in developing the populations. They acknowledge Texas A&M Institute for Genome Sciences and Society High Performance Compute Cluster (TIGSS HPCC) for the use of CheckMatrix. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture or University of California. The US Department of Agriculture is an equal opportunity provider and employer.

REFERENCES

1. Tanksley SD, Ganal MW, Prince JP, et al. High density molecular linkage maps of the tomato and potato genomes. *Genetics*. 1992;132:1141–1160.
2. Reinisch AJ, Dong JM, Brubaker CL, Stelly DM, Wendel JF, Paterson AH. A detailed RFLP map of cotton, *Gossypium hirsutum* × *Gossypium barbadense*: chromosome organization and evolution in a disomic polyploid genome. *Genetics*. 1994;138:829–847.
3. Ulloa M, Meredith WR Jr. Genetic linkage map and QTL analysis of agronomic and fiber quality traits in an intraspecific population. *J Cot Sci*. 2000;4:161–170.
4. Ulloa M, Meredith WR Jr, Shappley ZW, Kahler AL. RFLP genetic linkage maps from four $F_{2,3}$ populations and a consensus map of *Gossypium hirsutum* L. *Theor Appl Genet*. 2002;104:200–208.

5. Rong J, Abbey C, Bowers JE, et al. A 3347-locus genetic recombination map of sequence-tagged sites reveals features of genome organization, transmission and evolution of cotton (*Gossypium*). *Genetics*. 2004;166:389–417.
6. Lewin HA, Larkin DM, Pontius J, O'Brien SJ. Every genome sequence needs a good map. *Genome Res*. 2009;19:1925–1928.
7. Percival AE, Wendel JF, Stewart JM. Taxonomy and germplasm resources. In: Smith CW, Cothren JT, eds. *Cotton: Origin, History, Technology, and Production*. New York: Wiley-Blackwell; 1999:33–63.
8. Wendel JF, Cronn RC. Polyploidy and the evolutionary history of cotton. *Adv Agronomy*. 2003;78:139–186.
9. Ulloa M, Brubaker C, Chee P. Cotton. In: Kole C, ed. *Genome Mapping and Molecular Breeding in Plants: Technical Crops*. Vol 6. Berlin, New York, Tokyo: Springer;2007:1–49.
10. Grover CE, Zhu X, Grupp KK, et al. Molecular confirmation of species status for the allopolyploid cotton species: *Gossypium ekmanianum* Wittmack. *Genet Res Crop Evol*. 2014;62:103–114.
11. Wendel JF, Brubaker C, Alvarez I, Cronn R, Stewart JMcD. Evolution and natural history of the cotton genus. In: Paterson, AH, eds. *Genetics and Genomics of Cotton*. New York: Springer;2009:3–22.
12. Ulloa M, Abdurakhmonov IY, Perez-MC, Percy R, Stewart JMcD. Genetic diversity and population structure of cotton (*Gossypium* spp.) of the New World assessed by SSR markers. *Botany*. 2013;91:251–259.
13. Chen ZJ, Scheffler BE, Dennis E, et al. Toward sequencing cotton (*Gossypium*) genomes. *Plant Physiol*. 2007;145:1303–1310.
14. Endrizzi JE, Turcotte EL, Kohel RJ. Genetics cytology and evolution of *Gossypium*. In: Caspari EW, John G, eds. *Advances in Genetics*. San Diego, CA: Academic Press; 1985:271–375
15. Beasley JO. Meiotic chromosome behavior in species, species hybrids, haploids, and induced polyploids of *Gossypium*. *Genetics*. 1942;27:25–54.
16. Menzel MY, Brown MS. The significance of multivalent formation in three-species *Gossypium* hybrids. *Genetics*. 1954;39:546–557.
17. Brown MS. Identification of the chromosomes of *Gossypium hirsutum* L. by means of translocations. *J Hered*. 1980;71:266–274.
18. Guo WZ, Cai CP, Wang CB, et al. A microsatellite-based, gene-rich linkage map reveals genome structure, function and evolution in *Gossypium*. *Genetics*. 2007;176:527–541.
19. Yu Y, Yuan D, Liang S, et al. Genome structure of cotton revealed by a genome-wide SSR genetic map constructed from a BC_1 population between *Gossypium hirsutum* and *G. barbadense*. *BMC Genomics*. 2011;12:15. doi:10.1186/1471-2164-12-15.
20. Yu JZ, Kohel RJ, Fang DD, et al. A high-density simple sequence repeat and single nucleotide polymorphism genetic map of the tetraploid cotton genome. *G3*. 2012;2:43–58. doi:10.1534/g3.111.001552.
21. Hulse-Kemp AM, Lemm J, Plieske J, et al. Development of a 63K SNP array for cotton and high-density mapping of intraspecific and interspecific populations of *Gossypium* spp. *G3*. 2015;5:1187–1209. doi:10.1534/g3.115.018416.
22. Wang S, Chen J, Zhang W, et al. Sequence-based ultra-dense genetic and physical maps reveal structural variation of allopolyploid cotton genomes. *Genome Biology*. 2015;16:108. doi:10.1186/s13059-015-0678-1.
23. Frelichowski JE Jr, Palmer MB, Main D, et al. Cotton genome mapping with new microsatellites from Acala "Maxxa" BAC-ends. *Mol Genet Genomics*. 2006;275:479–491.
24. Blenda A, Scheffler J, Scheffler B, et al. CMD: a Cotton Microsatellite Database resource for *Gossypium* genomics. *BMC Genomics*. 2006;7:132–141.
25. Lacape JM, Jacobs J, Arioli T, et al. A new interspecific, *Gossypium hirsutum* × *G. barbadense*, RIL population: towards a unified consensus linkage map of tetraploid cotton. *Theor Appl Genet*. 2009;119:281–292.
26. Byers RL, Harker DB, Yourstone SM, Maughan PJ, Udall JA. Development and mapping of SNP assays in allotetraploid cotton. *Theor Appl Genet*. 2012;124:1201–1214. doi:10.1007/s00122-011-1780-8.
27. Gore MA, Fang DD, Poland JA, et al. Linkage map construction and quantitative trait locus analysis of agronomic and fiber quality traits in cotton. *The Plant Genome*. 2014;7:1–10.
28. Udall JA, Swanson JM, Haller K, et al. A global assembly of cotton ESTs. *Genome Res*. 2006;16:441–450.
29. An C, Saha S, Jenkins JN, Scheffler BE, Wikkins TA, Stelly DM. Transcriptome profiling, sequence characterization, and SNP-based chromosomal assignment of the EXPANSIN genes in cotton. *Mol Genet Genomics*. 2007;278:539–553.
30. Van Deynze A, Stoffel K, Lee M, et al. Sampling nucleotide diversity in cotton. *BMC Plant Biol*. 2009;9:125. doi:10.1186/1471-2229-9-125.
31. Lacape JM, Claverie M, Vidal RO, et al. Deep sequencing reveals differences in the transcriptional landscapes of fibers from two cultivated species of cotton. *PLoS ONE*. 2012;7:e48855. doi:10.1371/journal.pone.0048855.
32. Hulse-Kemp AM, Ashrafi H, Zheng X, et al. Development and bin mapping of gene-associated interspecific SNPs for cotton (*Gossypium hirsutum* L.) introgression breeding efforts. *BMC Genomics*. 2014;15:945. doi:10.1186/1471-2164-15-945.
33. Zhu QH, Spriggs A, Taylor JM, Llewellyn D, Wilson I. Transcriptome and complexity-reduced, DNA-based identification of intraspecies single-nucleotide polymorphisms in the polyploid *Gossypium hirsutum* L. *G3*. 2014;4:1893–1905. doi:10.1534/g3.114.012542.
34. Islam MS, Thyssen GN, Jenkins JN, Fang DD. Detection, validation and application of genotyping-by-sequencing based single nucleotide polymorphisms in upland cotton (*Gossypium hirsutum* L.). *The Plant Genome*. 2015;8:1–10.
35. Li FG, Fan GY, Wang KB, et al. Genome sequence of the cultivated cotton *Gossypium arboreum*. *Nat Genet*. 2014;46:567–572. doi:10.1038/Ng.2987.
36. Paterson AH, Wendel JF, Gundlach H, et al. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature*. 2012;492:423–427. doi:10.1038/nature11798.
37. Wang Z, Zhang D, Wang X, Tan X, Guo H, Paterson AH. A whole-genome DNA marker map for cotton based on the D-genome sequence of *Gossypium raimondii* L. *G3*. 2013;3:1759–1767.
38. Li F, Fan G, Lu C, et al. Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. *Nature Biotechnol*. 2015;33:524–530. doi:10.1038/nbt.3208.
39. Zhang T, Hu Y, Jiang W, et al. Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. TM-1) provides a resource for fiber improvement. *Nat Biotechnol*. 2015;33:531–537. doi:10.1038/nbt.3207.
40. Liu X, Zhao B, Zheng H-J, et al. *Gossypium barbadense* genome sequence provides insight into the evaluation of extra-long staple fiber and specialized metabolites. *Sci Rep*. 2015;5:14139. doi:10.1038/srep14139.
41. Ulloa M, Saha S, Jenkins JN, Meredith WR Jr, McCarty JC Jr, Stelly DM. Chromosomal assignment of RFLP linkage groups harboring important QTLs on an intraspecific cotton (*Gossypium hirsutum* L.) consensus map. *J Hered*. 2005;96:132–144.
42. Blenda A, Fang DD, Rami JF, Garsmeur O, Luo F, Lacape JM. A high density consensus genetic map of tetraploid cotton that integrates multiple component maps through molecular marker redundancy check. *PLoS ONE*. 2012;7:e45739. doi:10.1371/journal.pone.0045739.
43. Burke JJ. Moisture sensitivity of cotton pollen: an emasculation tool for hybrid production. *Agron J*. 2002;94:883–888.
44. Van Ooijen JW. *JoinMap® 4.1 Software for the Calculation of Genetic Linkage Maps in Experimental Populations*. Wageningen, The Netherlands: Kyazma BV; 2006.
45. Voorrips RE. MapChart: software for the graphical presentation of linkage maps and QTLs. *J Hered*. 2002;93:77–78.
46. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. *Nat Met*. 2012;9:357–359.
47. Shi Y, Li W, Li A, et al. Constructing a high-density linkage map for *Gossypium hirsutum* × *Gossypium barbadense* and identifying QTLs for lint percentage. *J Integr Plant Biol*. 2014;57:450–467.
48. Zhang K, Zhang J, Ma J, et al. Genetic mapping and quantitative trait locus analysis of fiber quality traits using a three-parent composite population in upland cotton (*Gossypium hirsutum* L.). *Molecular Breeding*. 2012;29:335–348.
49. Grover CE, Gallagher JP, Szadkowski Yoo MJ, Flagel LE, Wendel JF. Homoeolog expression bias and expression level dominance in allopolyploids. *New Phytol*. 2012;196:966–971.
50. Flagel LE, Wendel JF. Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. *New Phytol*. 2010;186:184–193.
51. Chelaifa H JA, Monnier A, Ainouche M. Transcriptomic changes following natural hybridization and allopolyploids in the salt marsh species *Spartina townsendii* and *Spartina anglica* (Poaceae). *New Phytol*. 2010;186:161–174.
52. Menzel MY, Richmond KL, Dougherty BJ. A chromosome translocation breakpoint map of *Gossypium hirsutum* genome. *J Hered*. 1985;76:406–414.