DEVELOPMENT OF MULTIPLEX AND SIMPLEX PLATFORMS FOR SNP-BASED INTROGRESSION OF A_2 AND D_1 GERMPLASM INTO COTTON *GOSSYPIUM HIRSUTUM* (L.)

A Thesis

by

AMMANI NAIDU KYANAM

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	David M. Stelly
Committee Members,	Jane Dever
	David Hawkins Byrne
Head of Department,	David Baltensperger

December 2016

Major Subject: Plant Breeding

ABSTRACT

Cultivated Upland cotton, *Gossypium hirsutum* L., is a partially diploidized allotetraploid species with relatively low levels of genetic diversity. Genetic gain through traditional breeding approaches is thus impeded. The wild species of the primary and secondary gene pools of cotton are approachable sources of agronomic traits of interest, but biological, cytogenetic, genetic and reproductive incompatibilities can impede progress. Genomic markers can alleviate certain difficulties, and expedite selective transfer of exotic species germplasm into one or more elite genotypes of a crop species. Coordinated development of Chromosome Segment Substitution Lines (CSSLs) using markers can in principle lead to complete representation of an alien genome in a cultivated crop.

Co-released with the CottonSNP63K Array, a BeadChip array for high-throughput genotyping of cotton, was a cluster file designed to facilitate automated genotype-calling germplasm from the primary genepool. Reported here is a new cluster file customized to germplasm from diploid species of the secondary gene pool. It significantly improves genotype call frequency and accuracy, and significantly increases the number of usable SNPs.

The first high-density interspecific genetic map of SNPs between cotton and diploid species was developed. It contains 14,411 SNPs, based on segregation in an A_2D_1-BC1F1 population from an interspecific cross of *G. hirsutum* and a A_2D_1 synthetic tetraploid. Genotypes of 72 BC1F1 plants were based on the CottonSNP63K array and

the new cluster file. Linkage analysis led to 26 linkage groups corresponding to the 26 chromosomes of cotton.

Utility of the CottonSNP63K Array and its associated SNPs for research and introgression breeding were significantly enhanced by demonstrating the derivation of simplex or low-plex SNP assays. A sample set of SNP mapped markers were validated on the KASP™ based assays and the conversion rate was estimated at 44%. This indicates a potential for development of ad hoc simplex SNP assays that can be applied to large populations for marker-assisted introgression, selection and down-stream breeding. KASP assays are applicable to DNA extracted non-destructively from seed or seedling at low-cost. Thus, validated CottonSNP63K SNPs can be used for targeted purposes, such as detection of rare recombination events or rare combinations of genes.

DEDICATION

I would like to dedicate this particular thesis to my family, Veeraswamy, Sai Kumari, Charitha and Haritha Kyanam who have loved me through the difficult person I was during the enjoyable struggle that was grad school.

ACKNOWLEDGEMENTS

I would like to acknowledge my deepest appreciation to my advisor, Dr. David Stelly, for providing guidance to think critically throughout my project. Thank you for providing me the opportunities and instilling in me the courage to present my research at several conferences. I also want to thank him for giving me the opportunity to interact with all his collaborators at different meetings. I would also like to acknowledge the distinguished faculty members who served as my committee members: Dr. Jane Dever and Dr. David Hawkins Byrne. They provided me valuable input and were very accommodating with their schedule. A special thanks to Dr. Wayne Smith and the department of Soil and Crop Sciences for giving me the opportunity to work as a teaching assistant for the plant breeding distance education program which not only provided me with financial support, but also gave me the opportunity to learn several useful skills.

I would like to thank Wayne Raska, Dr. Robert Vaughn and all the numerous student workers that have helped me care for my plants in the field and the greenhouse. I would also like to give special thanks to Dr. Amanda Hulse-Kemp for the countless hours she spent teaching me the right way to do different things and answer the numerous questions I have had. Also a special thanks to Dr. Robert Vaughn, who is very forgiving of all the terrible graduate students he has to deal with. I would also like to specially thank my colleagues at the Stelly lab: Andrea Maeda, Mariana Machado, Yu-

Ming Lin, Luis De Santiago and Dr. Bo Liu for their help with several research questions and their companionship during my term there.

As a part of my research, I used several facilities on the Texas A&M University campus. I would like to acknowledge the Agri-Genomics Laboratory and its lab managers: Dr. Fei Wang and Dr. Nithya Subramanian for teaching me DNA extractions and other genotyping methods used in my research. I would also like to acknowledge the Texas A&M Genome Science and Society (TIGSS), Dr. Penny Riggs, Dr. Claire Gill and Kelli Kochan for coordinating the SNP-array runs and allowing the use of their infrastructure during the preliminary stages of the cluster file development. Dr. Dirk Hays' lab at the Department of Soil and Crop Sciences allowed me to use their linkage mapping software that was used for most of the preliminary analyses. Dr. Silvano Ocheya and Dr. Trevis Huggins spent a portion of their time teaching me how to use the JoinMap® software and these lessons were crucial for the completion of my research work that contributed to the completion of this thesis. Ferdinand D'Souza, a dear friend, helped me with some crucial Excel tricks to make my data analysis more efficient.

I moved about 15,000 miles away from everything I knew to be normal, in the pursuit of a quality education and was for all intents without a family for the duration. I would like to acknowledge the people who have welcomed me into their homes during the holidays and other times when I was homesick. Amanda Holland Ray, LeAnn Hague, Dr. Amanda Hulse-Kemp, Andrea Maeda, Sabrina Allan Vaughn, Mariana Machado and Nancy Wahl, who have each, at different times, lent me a patient ear and advice during times of self-doubt I have experienced in the last three years. I appreciate

all the meals they have shared with me. I would also like to thank some of my other fellow graduate students: Alexandria Igwe, Brian Pfeiffer, Francisco Gomez, Dustin Wilkerson, Smit Dhakal, Laura Masor, Henry Awika, Yuanyuan Chen and so many others for their stress relieving conversations and for partying with me whenever I needed a break. I would also like to thank my most amazing roommate, Bara Safarova.

I would like to acknowledge my best friends from back home, Lakshmi Mupparthi, Sai Kiran Janaki, Jyothi Boinapally and Roopa Gandi who have each made sure to stay in touch with me in spite of the physical distance and time difference. I would also like to thank my sister, Charitha and her husband, Vishnu Pavan for becoming my *ipso facto* cool parents in the west. I would also like to thank my other sister, Haritha for being the one to convince me to attend Texas A&M University and the tremendous amount of guidance given throughout my life. I also want to thank my sweetest nephews, Tanai and Nivant, for the magical smiles and hugs they save for when their little aunt visits. Finally, I want to thank my parents, Sai Kumari and Veeraswamy Naidu for their encouragement and support, both financial and emotional. I want to specially thank them for letting me out of the protective nest and believing that I would land on my feet.

Through the journey at Texas A&M and the duration of my stay in the States, I have had the opportunity to meet several talented and interesting individuals who have positively influenced my growth as an individual and have made me a better person. If I have missed mentioning YOU by name, this is me saying “THANK YOU”.

NOMENCLATURE

AFLP	Amplified Fragment Length Polymorphism
BC1F1	Backcross 1 Filial Generation 1
cM	centiMorgan
CSSL	Chromosome Segment Substitution Line
IL	Introgression Line
KASP	(K)Competitive Allele Specific PCR
LG	Linkage Group
LOD	Logarithm of Odds
MAS	Marker Assisted Selection
NIL	Near-Isogenic Line
NPGS	National Plant Germplasm System
PCR	Polymerase Chain Reaction
QTL	Quantitative Trait Locus
RFLP	Restriction Fragment Length Polymorphism
SSR	Simple Sequence Repeat
SNP	Single Nucleotide Polymorphism
USDA	United States Department of Agriculture

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION AND LITERATURE REVIEW

Importance of Cotton

Cotton is the world's most important natural textile fiber crop. It is the basic component for thousands of consumer and industrial products manufactured in the United States and throughout the world. Furthermore the contributions made by cotton to the textile, feed, food, and other industries continue to grow in importance. In 2015, in the U.S., cotton was grown on over 8.6 million acres (~3.5 M ha.) which produced about 12.9 million bales of cotton. Business revenue stimulated by the crop in the U.S. economy is estimated at over $75 billion (National Cotton Council, www.cotton.org). The U.S., India and China collectively provide two-thirds of the world's cotton production. (www.usda.gov). The U.S. is the 3rd largest producer and leading exporter with over 10.5 million bales contributed to the world cotton exports, accounting for over 37% of the total export market of raw cotton (The National Cotton Council, www.cotton.org). The U.S. textile manufacturers use an annual average of 7.6 million bales of cotton, majority (93%) of which contributed to the making of apparel and home furnishings. A small percentage (7%) is used for manufacturing industrial products. Furthermore, two thirds of the 6.5 billion tons of cottonseed produced is used as animal feed and the remaining is used in the production of cottonseed oil used in the manufacture of margarines, cooking oils, salad dressings etc. In addition to being a

significant contributor to the economies of developing countries like Uzbekistan and Pakistan through export revenue, cotton farming provides both direct and indirect employment to many millions of people in the world's poorest countries (Ergon, 2008).

The genus *Gossypium* comprises 52 species that presently include 7 known tetraploid 2n=4x=52 species and 45 diploid 2n=2x=26 species (Jonathan F. Wendel & Grover, 2015), where n represents the haploid number of chromosomes. The diploid species are classified into eight genomic groups (A-G and K). The genome designations are based on collective observation of pairing behavior, chromosome size, relative fertility in interspecific hybrids, and adaptation to different regions (Beasley, 1942; Jonathan F Wendel, Brubaker, Alvarez, Cronn, & Stewart, 2009). The African-Arabian species include the lint-bearing A-genome species, also known as the Old World cottons, as well as the B-, E- and F-genome species. The New World cottons that originated in Northern and Central America include the D-genome species. The Australian species include the diploids with the C, G and K genomes.

The ancestral divergence of cotton A- versus D-genome lineages is estimated to have occurred approximately 5 – 10 million years ago (mya). About 1-2 mya, a single polyploidization event involving an A-like genome and a D-like genome led to the formation of the allotetraploid $[AD]_n$ ancestor that subsequently diverged into the present day tetraploid species (Jonathan F Wendel & Cronn, 2003). *G. raimondii* Ulbrich [D_5] has been determined to be most closely related species to the D-genome extant contributor, while *G.herbaceum* L. [A_2] is the most closely related species to the A-genome extant contributor, of the tetraploid genome respectively (Jonathan F Wendel &

Cronn, 2003). The tetraploid species are *G. hirsutum* L. $[AD]_1$, *G. barbadense* L. $[AD]_2$, *G. tomentosum* Nuttall ex Seemann $[AD]_3$, *G. mustelinum* Miers ex Watt$[AD]_4$, *G. darwinii* Watt $[AD]_5$, *G. ekmamianum* Wittmack $[AD]_6$ (C. E. Grover et al., 2015) and $[AD]_7$ (Wendel – unpublished). The genome designations and their geographic distributions lend themselves to facile description of three distinct genepools (Harlan & de Wet, 1971) (**Table 1**). The primary pool usually includes the cultivated species, and in the case of *Gossypium*, it includes all the tetraploid species. Interspecific crosses between the species in this pool result in the formation of fertile hybrids. The secondary genepool includes the D-genome American species and the African-Arabian diploid species with genome groups of A, B, and F. The crosses between the species in this gene pool to *G. hirsutum* and the other species in the primary gene pool are possible, but challenging. The tertiary genepool includes the E genome African-Arabian species and all the Australian species that have the genome designations of C, G and K. The crosses between the primary and the tertiary gene pool are usually anomalous, sterile or lethal.

Table 1: List of *Gossypium* species classified into gene pools (NPGS, 1997).

Germplasm Pool	Species Included			
Primary Germplasm Pool	*G. hirsutum* $(AD)_1$	*G. barbadense* $(AD)_2$	*G. tomentosum* $(AD)_3$	*G. mustelinum* $(AD)_4$
	G. darwinii $(AD)_5$	*G. ekmamianum* $(AD)_6$	$(AD)_7$ (Unpublished)	
Secondary Germplasm Pool	*G. herbaceum* (A_1)	*G. arboreum* (A_2)	*G. anomalum* (B_1)	*G. triphyllum* (B_2)
	G. capitis-viridis (B_3)	*G. trifurcatum* (B)	*G. longicalyx* (F_1)	*G. thurberi* (D_1)
	G. armourianum $(D_{2\text{-}1})$	*G. harknessii* $(D_{2\text{-}2})$	*G. davidsonii* $(D_{3\text{-}d})$	*G. klotschianum* $(D_{3\text{-}k})$
	G. aridum (D_4)	*G. raimondii* (D_5)	*G. gossypioides* (D_6)	*G. lobatum* (D_7)
	G. trilobum (D_8)	*G. laxum* (D_9)	*G. turneri* (D_{10})	*G. schwendimanii* (D_{11})
Tertiary Germplasm Pool	*G. sturtianum* (C_1)	*G. robinsonii* (C_2)	*G. bickii* (G_1)	*G. australe* (G)
	G. nelsonii (G)	*G. costulatum* (K)	*G. cunninghamii* (K)	*G. enthyle* (K)
	G. exgiuum (K)	*G. londonerriense* (K)	*G. marchantii* (K)	*G. nobile* (K)
	G. pilosum (K)	*G. populifolium* (K)	*G. pulchellum* (K)	*G. rotundifolium* (K)
	G. anapoides (K)	*G. stocksii* (E1)	*G. somalense* (E_2)	*G. areysianum* (E_3)
	G. incanum (E4)	*G. bricchettii* (E)	*G. benadirense* (E)	*G. vollensenii* (E)
The genomic grouping of Australian species is under study. Where used, () indicate provisional genomic placement for the species in question				

The lint-bearing species from which cultivated species were domesticated are restricted to the A-genome and AD-genome species (Stephens, 1947) with an exception of the Madagascan species, *Gossypium brevilanatum*. Evidence suggests that prior to domestication each of the four cultivated species existed as distinct wild species (Brubaker & Wendel, 1994). The divergent, geographically isolated ancestors of cultivated cotton were independently subjected to convergent domestication by ancient

human cultures in both the Old and New Worlds. This parallel domestication process involved four species, two from Americas or the New World species, *G. hirsutum* and *G. barbadense*, and two from Africa–Asia or Old World species, *G. arboreum* and *G. herbaceum*. Cultivation of these four species still contributes to the world cotton production, but *G. hirsutum* L., also known as Upland cotton, is grown in over 40 nations across tropical and temperate regions of the world. Consequently, over 95%, of the commercial cotton produced is Upland cotton (Lin et al., 2005), thereby making it the most economically important species. The remaining ~5% of cotton produced is mostly from *G. barbadense* L., such as Pima cotton, with meager contributions from the diploid cottons *G. arboreum* L. and *G. herbaceum* L. grown in parts of Asia.

Need for Introgression

Gossypium hirsutum L. is a highly diverse species. As per most widely accepted classification (J. B. Hutchinson, Silow, R. A., Stephens, S. G.,, 1947), the three taxonomic varieties are recognized that correspond to races "*latifolium*", "*punctatum*", and "*marie-galante*". Evidence shows that after initial domestication of Upland cotton in the Yucatan peninsula (Stephens, 1958), the original cultivars were widely dispersed. These later developed into localized derivatives, the most important of which was the annualized race "*latifolium*" in the Mexican stock. The "*latifolium"* derivative, after agronomic improvement, spread throughout Mesoamerica through human-mediated dispersal. Molecular evidence show that the modern annual forms of *G. hirsutum* L. grown in the United States traces back to those Mexican Stocks (Van Esbroeck &

Bowman, 1998; Jonathan F Wendel, Brubaker, & Percival, 1992). As a genetic consequence of domestication, multiple lines of evidence indicate that this species of cotton that currently dominates that world cotton production have inherently low levels of diversity (Fang et al., 2013; Iqbal, Reddy, El-Zik, & Pepper, 2001; Jonathan F Wendel et al., 1992). For example, a comparative study of the nucleotide diversity in *G. hirsutum* for the homeologous *Adh* locus, responsible for anaerobic respiration, to other *Gossypium* spp. showed that the measure of observed heterozygosity at the *Adh*A locus was zero in most cases. The hypothesized reasons for such low levels of reduced nucleotide diversity can be attributed to cultivated cotton's recent polyploidization, a low mutation rate and its self-pollinating reproductive biology (Small, Ryburn, & Wendel, 1999).

Genetic variation for agriculturally important traits to improve yield potential, disease resistance, stress resilience, water-use efficiency and adaptability to changing climatic conditions are unlikely to be found within the limited diversity in the elite lines of the current cultivars. This leaves the breeders to explore alternatives to enable genetic improvement like introgression, mutagenesis and different methods of genetic engineering. The exotic germplasm pools of *Gossypium* include several species with potentially useful traits. Vavilov was the first to report the use of crop relatives as potential new sources of genes for improving agricultural productivity (Vavilov, 1940). This potential, along with the National Academy of Sciences report after the Southern Corn Blight disaster, fueled the establishment of the International Plant Genetic Resources Institution (IPGRI) (Tanksley & McCouch, 1997). The IPGRI is currently

known as the Bioversity International and is part of the 15 research centers working under the umbrella of the Consultative Group on International Agricultural Research (CGIAR) Consortium. This institution was entrusted with coordinating the efforts in the collection and preservation of plant germplasm materials. As of 2015, an expenditure of $36.9 million was incurred in maintaining the germplasm reserves at Bioversity (Bioversity, 2015). *Ex situ* conservation represents the most significant means of conservation where the accessions are kept in specialized facilities known as genebanks. In 1994, the CGIAR centers signed a treaty with the Food and Agriculture Organization of the United Nations, bringing their collections into the International Network of *Ex situ* Collections (International Treaty on Plant Genetic Resources for Food and Agriculture). Currently, there are over 1,750 individual genebanks holding about 7.4 million accessions in total (FAO, 2010).

Traditional breeding programs involving exotic germplasm are considered time-consuming and labor intensive. Though laden with challenges that often constrain usage to simply inherited traits, the contributions of wild germplasm through wide-cross breeding have been profound. Using wild germplasm to transfer genes associated with disease resistance is one of the most popular applications of this technique. For example, the rust resistance gene (*Lr*19) from *Agropyron elongatum* in commercial wheat (Hoisington et al., 1999) and genes for resistance to brown plant hopper, bacterial blight and blast in rice have been introgessed from wild species (Brar & Khush, 1997). Examples of the transfer of more complex inheritance pattern QTLs associated with

improved fruit quality were performed in tomato (Rick, 1974; Ronen, Carmel-Goren, Zamir, & Hirschberg, 2000).

A_2D_1 Synthetic Tetraploid

Development and cytological analysis of interspecific hybrids were key to the elucidation of genomic diversification among 26-chromosome cotton species and their relationship to each other as well as to the 52-chromosome cotton species (Beasley, 1942). Experiments with interspecific hybrids of Asiatic and new world diploid cottons at the Texas Agricultural Experiment Station (Menzel & Brown, 1954) found synthetic hybrids could overcome the ploidy difference between the Upland cotton and exotic cotton species that impede trait transfer. Popularization of the use of interspecific hybrids for introgression of traits from diploids in cotton led to the development of a collection of interspecific lines. Four decades of research utilizing interspecific hybrids, especially at the Gembloux Agricultural University, Belgium, led to the development of a large collection of several kinds of interspecific hybrids. This collection includes 21 bispecific diploids, 12 bispecific triploids, 6 bispecific synthetic allotetraploids, 8 trispecific synthetic allotetraploids, 11 bispecific synthetic allohexaploids, 11 bispecific pentaploids and 13 monosomic alien addition lines involving species of A, B, C, D, E, F, and G genomes of genus *Gossypium* (Mergeai, 2006). Some notable examples of trait introgression in cotton through interspecific introgression are improved fiber strength from *G. thurberi* (Stewart & Hsu, 1977), increased gossypol content from *G. raimondii* – *G. thurberi* bispecific hybrid (Rhyne & Smith, 1965) and non-glabrous leaves from *G.*

armorianum (Meyer, 1957) that both foster resistance to *Heliothis* spp., resistance to bacterial blight caused by *Xanthomons malvacearum* (Innes, 1966) and resistance to rust caused by *Puccinia cacabata* (Percival & Kohel, 1990). QTL mapping of yield-associated traits indicated that wild species are more than just a source of quantitative traits (Moncada et al., 2001).

A_2D_1 is one such synthetic allotetraploid derived from an interspecific cross between *Gossypium arboreum* L. and *Gossypium thurberi* Tod. This hybrid was developed by crossing *G. arboreum* L. with *G. thurberi* Tod. followed by chromosome doubling using colchicine treatment.

Gossypium arboreum

The wild progenitor of *G. arboreum* L. is unknown and there is speculation about its origin. There are two possible geographical regions for domestication of this species. In the first location, Madagascar, there are two primitive forms of *G. arboreum*. The arborescent form found in the xerophytic woodlands and the primitive cultigen found in association with human settlements (J Hutchinson & Dalziel, 1954). The Indus Valley Civilization is the second possible location because archeological evidence suggests that it is the place where cotton was first cultivated (Gulati & Turner, 1929). The race "*indicum*" represent the most primitive form of *G. arboreum* and the cultivated forms are thought to be the annualized derivatives of this perennial form. Their morphology varies from multi-branched shrub with scanty coarse lint to unbranched subshrubs with higher-quality lint. Morphological and genetic similarities (Silow, 1944) between *G.*

arboreum race *indicum* and *G. herbaceum* race *acerifolium* cultivars has been cited as evidence of progenitor-derivative relationship (Joseph Hutchinson, 1962; J Hutchinson & Dalziel, 1954). Molecular data supports the alternative hypothesis that the two species were independently domesticated from divergent wild progenitors. These observations show that morphological similarities can be attributed to parallel retention of characteristics from a common ancestor or from post-domestication introgression between *G. arboreum* and *G. herbaceum* as they came in contact along the Indian Ocean trade routes (Jonathan F Wendel, 1989). This species was used as the A-genome contributor to the development of the A_2D_1 synthetic allotetraploid.

Gossypium thurberi

G. thurberi Tod. is the only wild diploid species that occurs naturally in the United States. It is found distributed from the state of Arizona to the state of Sonora, Mexico (Fryxell, 1976). The species can be described as a small tree or a perennial shrub that is around 3m tall. The species have leaves that are smooth, palmately lobed with flowers that are white to yellow in color. The mature bolls are trilocular and seeds are blackish in color with no fiber (Ulloa, 2014). The species is believed to possess some cold hardiness, escaping frost damage mostly through defoliation. It is a D-genome species and the second diploid species contributor to the synthetic allotetraploid A_2D_1 that was used by Beasley to introgress fiber strength into *G. hirsutum* (Beasley, 1942).

Breeding efforts with A_2D_1 Synthetic

In early 1970's, research work utilizing the A_2D_1 synthetic tetraploid was initiated at the Southern Region, Agricultural Research Station, USDA, Pee Dee Experiment Station, Florence, S. C. for improving fiber strength of medium staple Upland cotton (Culp & Harrell, 1973). Under the pedigree method of breeding, they identified two lines, Earlistaple-7 and Pee Dee 4381-54 with improved fiber qualities. Due to lower recovery rates through pedigree breeding method, backcrossing was explored as an additional method, which led to the development of a valuable breeding line (Q). Some of the challenges faced by the group were small population sizes and lack of means to select desirable/promising recombinants in early generations. They recommend selection for desirable traits in early generations and the use of systems that increases hybridizations that give desirable recombinants. Early generation selections can also facilitate the elimination of undesirable traits associations and eliminate the need for large population sizes. Wide-cross programs such as these are also impeded by numerous reproductive and biological impediments including F1 hybrid sterility, reduced recombination, and linkage drag.

Tools Needed for Establishment of Introgression Libraries

Due to the challenges associated with trait introgression from the wild species of cotton, the establishment of curated collections of accessions of wild relatives of crop species that can be made available to breeding programs for research purposes are inadequate (Tanksley & McCouch, 1997). It is vital that tools be developed that would

allow breeders to rapidly use the genetic potential these wild species offer. Introduction of different molecular marker technologies and the associated marker-assisted selection strategies have improved the efficacy of plant breeding programs (Eathington, Crosbie, Edwards, Reiter, & Bull, 2007). RFLP's were widely used in early stages of molecular marker development (Lander & Botstein, 1989). The invention of PCR technology replaced the low-throughput RFLP markers with RAPD, AFLP, and SSR markers. Due to low levels of reproducibility of RAPDs and cumbersome detection methodology of AFLPs, neither was applicable for molecular breeding. SSRs or microsatellite DNA markers that eliminated the aforementioned drawback led them to become the most widely used markers in the early 21st century. However, SNPs markers have proven to be the more abundant than SSRs in a genome. The rapid development of a diverse range of SNP genotyping methods signify their importance, applicability and flexibility. While the rates and degrees of polymorphism of individual SNPs are lower than those of an SSR, their abundance and amenability to high- or ultra-high throughput with automation more than compensates for this drawback (Mammadov, Aggarwal, Buyyarapu, & Kumpatla, 2012).

Advances in molecular marker technologies have strengthened breeding programs that allowed for tracking the wild alleles and monitoring complete wild genome representation in an elite background. This led to the development of introgression line libraries (Zamir, 2001). A library of lines in which each line contains one or more defined chromosome segments that originate from the exotic species in an otherwise uniform elite genetic background are known as an Introgression Lines (ILs).

The unique feature of the ILs is that they are a permanent seed source that can be maintained by simple selfing methodology. This facilitates rapid re-screening of particular interspecific crosses that would otherwise be challenging without the permanent lines due to the time-constraints involved in recreating advanced backcross generations of the desired interspecific cross.

Introgression Lines have other applications in different breeding programs. The lines in an introgression library are an efficient tool for mapping agronomically important traits. Since they only differ from the cultivated lines by one or few defined chromosome segments, they reduce sterility issues associated in making wide-crosses. They are a permanent resource and can be tested by several different groups, in different locations and at different times; their phenotypic data can also be collated and made available to researchers worldwide. The lines are homozygous for the introgression; they can be used to develop line by tester cross that could help in understanding effects of heterozygosity on phenotypes and lead to the identification of genes that show heterosis. Recombination-mediated reduction of the QTL-carrying segments helps in fine mapping for these QTLs. Such lines with overlapping segments can contribute to efforts in breaking linkage drag. Crosses between lines containing individual QTLs can contribute to closer examination of phenotypic effects of QTL interactions and may help understand epistasis.

The earliest known use of an introgression library was the whole-chromosome substitution introgression lines used in the analysis of complex traits in common wheat (Kuspira & Unrau, 1957). Segmental introgression line development, aided by molecular

markers, was later pioneered in tomato. In *Solanum lycopersicum* (previously *Lycopersicon esculentum*) 50 ILs were developed, each containing a single introgressed segment of *S. pennellii* (previously *L. pennellii*); 350 RFLP markers were used track whole alien genome representation within the lines (Yuval Eshed & Zamir, 1994). This population was used to first identify yield-associated QTLs (Y. Eshed & Zamir, 1995), and also analyze epistatic and environmental interactions. This subsequently led to high-resolution mapping of the trait (Fridman, Pleban, & Zamir, 2000). Development of such populations was challenging when molecular markers were in the nascent stages of development. However, advances in molecular markers have considerably reduced the number of generations required for developing similar populations (Young, 1999). These successes led to marker-based efforts directed at the development of similar libraries for other agronomically important crops. Sequencing efforts in different crops will facilitate the development of informative markers that have the potential to be used in breaking linkage drag associated with wide-crosses

Though known by different names such as Chromosome Segment Substitution Lines (CSSLs) in rice (Kubo et al., 2002), or Near Isogenic Lines (NILs) or QTL-NILs in barley (Von Korff, Wang, Léon, & Pillen, 2004), or Backcross Inbred Lines (BILs) in lettuce (Jeuken & Lindhout, 2004), the methodology used in their construction is similar. In cotton, CSSLs were developed using *Gossypium barbadense* in the elite background of TM-1, the genetic standard of *G. hirsutum* that includes a total of 330 individual lines (P. Wang, Ding, Lu, Guo, & Zhang, 2008). Modification of drought related physiological traits were achieved using NILs developed using marker-assisted selection

from crosses between *G. barbadense* and *G. hirsutum* (Levi et al., 2009). Similarly, a total of 28 QTLs for fiber quality, including four for fiber elongation, eight for fiber fineness, four for fiber strength, four for fiber length, six for fiber uniformity, one for boll weight, and one for boll number were identified in backcross-inbred families developed from a cross between *G. hirsutum* and the Hawaiian cotton, *G. tomentosum* (Zhang et al., 2011).

The volume of wild genes that come through the traditional backcross breeding methods can be problematic, especially in early backcross generations due to the fact that the introgressed segments are very large. Linkage drag is then more likely to exist between a favorable gene and an unfavorable gene. One can infer that the extent of linkage drag will increase in the remaining heterozygous segments during traditional backcrosses if there is suppression of recombination in the homoelogous regions. This inference is based on the reports of favored recombination seen in the homologous regions of the chromosome in crops like tomato, rice and cotton (Jia, Jia, Wang, & Liu, 2012; Zheng et al., 2016). Due to this suppression, it is challenging to break the unfavorable linkages within introgressed segments. A breeding method to overcome such situations is to make early generation selection for nearby crossover products. This can be followed by intercrossing between selected pairs of recombinants that minimize homologous recombination such as intercrosses between two complementary introgression lines that overlap only in the region containing the gene of interest. This results in an F1 where the region of "homology" (homozygous for alien segments or heterozygous with "bridge" genome) is between the wild introgression segments with

the gene of interest and this encourages recombination in that common region and therefore increases the chances of breaking the unfavorable linkages (Zheng et al., 2016). The limitation of this method is that favorable crossover products on both sides of the target gene must be identified very early in the backcross, ideally in the BC1F1 generation.

The CottonSNP63K array-based Illumina Infinium II SNP genotyping assays was developed by the Stelly lab at the Texas A&M University, UC Davis, CSIRO of Australia and other members of the International Cotton Consortium (Hulse-Kemp et al., 2015). This chip was designed with 70,000 SNP assays, 50,000 (~70%) for intraspecific SNPs and 20,000 (~30%) interspecific. Approximately 90% were synthesized successfully, yielding 45,104 putative intraspecific assays and 17,954 putative interspecific assays. This multiplexed genotyping platform is high-throughput, easy to use, allows for construction of high-density maps and has a wide range of potential uses for applied breeding and breeding research. Such arrays can expedite high-quality diversity analyses, the development of high-density linkage map, localization of quantitative trait loci (QTLs), all of which can then be used to design efficient marker assisted selection studies. However, for targeted backcross breeding for introgression line development, the highly multiplexed arrays are relatively expensive. The cost would skyrocket for introgression effort requiring the recovery recombinants i.e., when large numbers of seed or seedlings must be screened (Zheng et al., 2015). The availability of a high-density linkage map and an efficient, simplex or low-plex genotyping pipeline that enables affordable genotyping in a targeted segment (or segments) of interest, can

facilitate introgression breeding and in the pyramiding of genes or traits of interest. In addition to providing a high-throughput genotyping platform, the SNP marker sequences included in the Cotton SNP array for the 63,000 SNP markers have high potential to be converted to PCR-based assays. In the preliminary analysis with a small sample set of individuals in the mapping population, it was found that approximately 17,000 of these markers are polymorphic for the A_2D_1- BC1F1 population. This indicated that this genotyping platform would be suitable to genotype the rest of the individuals in the mapping population, and subsequently develop a high-density linkage map. Concurrently, a sample set of the mapped markers will be selected to evaluate the conversion rate at which the array markers can be converted to a PCR-based KASP™ assay.

KASP™ assays are a proprietary SNP genotyping system from LGC (http://www.lgcgroup.com/). Their unique labeling mechanism, which allows for visualization of the genotypes, makes the system comparatively flexible and inexpensive. When combined with high-throughput DNA extraction method (Zheng et al., 2015), it provides cotton breeders with a very powerful tool that will enable the introgression of traits of interest and MAS for rare recombination events or rare multigene combinations.

The ultimate goal of this project is the development of tools that would be valuable in the construction of Chromosome Segment Substitution Lines (CSSLs) for the A_2D_1 synthetic and thereby will provide cotton breeders with tools to expedite trait introgression from these two diploid cotton species. The objective of this project to

develop [1] a finished cluster file that is customized for A_2D_1 introgressed populations, [2] a BC1F1 mapping population, [3] a high-density linkage map or bin map of markers for the BC1F1 population of the A_2D_1 synthetic tetraploid, [4] a number of functional KASP assays for position defined A_2D_1-(AD)1 SNPs, [5] an estimate of conversion rate of SNP marker sequences in the CottonSNP63K cluster file into simplex KASP™ assays, and [6] an advanced BC4F1 population. These resources can be used to generate Chromosome Segment Substitution Lines (CSSLs) for A_2D_1 synthetic tetraploid in *G. hirsutum* background and to rapidly expand the number of useful KASP™ assays.

CHAPTER II

INTEGRATED PLATFORM FOR SNP-BASED INTROGRESSION OF A_2 AND D_1 GERMPLASM INTO UPLAND COTTON *GOSSYPIUM HIRSUTUM* (L.) AND ITS ANALYSIS

Introduction

Vavilov was the first to report the use of crop relatives as potential new sources of genes for improving agricultural productivity (Vavilov, 1940). This potential, along with the National Academy of Sciences report after the Southern Corn Blight disaster, fueled the establishment of the International Plant Genetic Resources Institution (IPGRI) (Tanksley & McCouch, 1997). The IPGRI is currently known as the Bioversity International and is part of the 15 research centers working under the umbrella of the Consultative Group on International Agricultural Research (CGIAR) Consortium. This institution was entrusted with coordinating the efforts in the collection and preservation of plant germplasm materials. As of 2015, an expenditure of $36.9 million was incurred in maintaining the germplasm reserves at Bioversity (Bioversity, 2015). *Ex situ* conservation represents the most significant means of conservation where the accessions are kept in specialized facilities known as genebanks. In 1994, the CGIAR centers signed a treaty with the Food and Agriculture Organization of the United Nations, bringing their collections into the International Network of *Ex situ* Collections (International Treaty on Plant Genetic Resources for Food and Agriculture). Currently,

there are over 1,750 individual genebanks holding about 7.4 million accessions in total (FAO, 2010). The development and curation of collections of wild relatives of crop species facilitates the evaluation and usage of wild germplasm in crop breeding programs. However, there are numerous biological, genetic and genomic factors that can impede or preclude hybridization, introgression and use of wild germplasm (Hadley & Openshaw, 1980).

Advances in molecular marker technologies have enabled breeding programs to indirectly track wild germplasm versus cultivated germplasm. The ability to use markers to monitor wild genome representation in interspecific breeding materials, e.g., during backcross inbred development, enabled development of introgression line libraries (Zamir, 2001), where each introgression line (IL) contains one or more defined chromosome segments that originate from the exotic species in an otherwise uniform elite genetic background. The most valuable feature of the ILs may be that they provide a permanent seed source for each genotype that can be maintained by simple selfing methodology. This facilitates extensive phenotyping, e.g., replicated testing, multi-location, multi-year and multi-trait evaluations of each line from a particular interspecific cross. As permanent lines, ILs remove some major experimental time-constraints, e.g., the need to recreate advanced backcross generations of the desired interspecific cross for multiple evaluation experiments. Furthermore, an introgression library is an efficient tool for mapping agronomically important traits, because each IL differs from the cultivated recurrent parent only by one or few marker-defined chromosome segments. This composition minimizes sterility issues that often hamper

analysis of F2, F3 and RIL wide-cross populations, for which half of the germplasm is alien and segregating widely. Since ILs are homozygous for introgressed segments, they can be used to develop line-by-tester testcrosses to help in understanding effects of heterozygosity on phenotypes and lead to identification of genes that show heterosis. Recombination-mediated reduction of the QTL-carrying segments helps in fine-mapping of QTLs. ILs with overlapping donor segments can contribute to efforts in breaking linkage drag. Crosses between lines containing different QTLs can reveal phenotypic effects of QTL interactions and epistasis (Zamir, 2001). The development of segmental ILs by molecular marker assisted selection was pioneered in tomato (Yuval Eshed & Zamir, 1994). This population was used to first identify yield-associated QTLs (Y. Eshed & Zamir, 1995), and also to analyze epistatic and environmental interactions. This subsequently led to high-resolution mapping of the trait (Fridman et al., 2000). These successes led to marker-based efforts directed at the development of similar libraries for other agronomically important crops. Though known as Chromosome Segment Substitution Lines (CSSLs) in rice (Kubo et al., 2002), or Near Isogenic Lines (NILs) or QTL-NILs in barley (Von Korff et al., 2004), or Backcross Inbred Lines (BILs) in lettuce (Jeuken & Lindhout, 2004), the methodology used in their construction is similar.

Cotton (*Gossypium* spp.) is the world's most important natural fiber crop. The genus *Gossypium* comprises of 52 species that presently include 7 known tetraploid (2n=4x=52) (C. Grover et al., 2015; Jonathan F Wendel et al., 2009) and 45 diploid (2n=2x=26) species, where n represents the haploid number of chromosomes. The eight diploid genome group designations (A-G and K) are based on collective observation of

pairing behaviour, chromosome size, and relative fertility in interspecific hybrids (Beasley, 1942; Jonathan F Wendel et al., 2009). A single polyploidization event between an A-like genome and the D-like genome led to the formation of the allotetraploid $[AD]_n$ ancestor that diverged into the present day tetraploid species (Jonathan F Wendel & Cronn, 2003). *Gossypium hirsutum* L. $[AD]_1$, also known as Upland cotton, is the most economically important species as it is grown in over 40 nations across tropical and temperate regions of the world. Consequently, 93% of the commercial cotton produced is Upland cotton (Lin et al., 2005).

Multiple lines of evidence (Fang et al., 2013; Iqbal et al., 2001) indicate levels of genetic diversity in *G. hirsutum* are low, and that they can be attributed to its recent polyploidization and self-fertilizing reproductive biology (Small et al., 1999; Weeden & Wendel, 1989). This low diversity impedes genetic improvement and necessitates the augmentation of natural diversity by other means, e.g., mutagenesis, biotechnology, and introgression from other species included in the genus. Though the time-consuming nature of introgression breeding often constrains the use of exotic species to simply inherited traits, the contributions of wild germplasm through wide-cross breeding have been profound. Using wild germplasm to transfer genes associated with disease resistance is one of the most popular application of this technique in different crops (Brar & Khush, 1997; Hoisington et al., 1999). In order to facilitate introgression while reducing linkage drag, 330 CSSLs were developed using a *G. barbadense* L. donor, with segments introgressed into the elite background of TM-1, the genetic standard of *G. hirsutum* (Levi et al., 2009; P. Wang et al., 2008; Zhang et al., 2011). These lines were

used to characterize several drought-related physiological traits and their relationships (Levi et al., 2009). Similarly, a total of 28 QTLs for fiber quality were identified in backcross-inbred families developed from a cross between *G. hirsutum* L. and the Hawaiian cotton, *G. tomentosum* (Zhang et al., 2011). It would be beneficial to have similar CSSLs for other species from the exotic germplasm pool, because diploid species with traits of interest in the secondary genepool (Harlan & de Wet, 1971), like *G. arboreum* L. and *G. thurberi* Tod. cannot be directly crossed with the tetraploid species and used to aid breeding efforts for these species. Interspecific hybrids of Asiatic and New World diploid cottons at the Texas Agricultural Experiment Station (Menzel & Brown, 1954) found synthetic hybrids could overcome the ploidy difference between the Upland cotton and exotic cotton species and allow the formation of fertile progenies from these crosses. Researchers in Gembloux Agricultural University, Belgium, developed a large collection of interspecific hybrids lines involving species of A, D, B, C, E, F, and G genomes of cotton (Mergeai, 2006). Notable examples of trait introgression in cotton using interspecific introgression are improved fiber strength from *G. thurberi* (Stewart & Hsu, 1977), increased gossypol content from *G. raimondii – G. thurberi* bispecific hybrid (Rhyne & Smith, 1965) and non-glabrous leaves from *G. armourianum* (Meyer, 1957). QTL mapping of yield-associated traits indicated that wild species are more than just source of quantitative traits (Moncada et al., 2001). We have developed a BC1F1 population that is derived from a cross between TM-1, the genetic standard for *G. hirsutum* L. and a synthetic allotetraploid, $2[A_2D_1]$ that was derived from an interspecific cross between *G. arboreum* L. and *G. thurberi* Tod. This hybrid has

been previously used for improving fiber strength of medium staple Upland cotton (Culp & Harrell, 1973) at the Pee Dee Experiment Station, Florence, S. C. However, progress was hampered by the lack of means to select desirable recombinants at that time. The advent of IL breeding strategies and newly enhanced simplex and highly multiplexed SNP genotyping platforms creates an opportunity to revisit previous A_2D_1 breeding efforts. Using new molecular tools, it may be possible to markedly enhance A_2 and D1 genome introgression work at both analytical and applied breeding levels.

The CottonSNP63K array is a highly multiplex SNP genotyping tool developed by our laboratory and the International Cotton Consortium that can yield up to 45,104 putative intraspecific assays and 17,954 putative interspecific assays (Hulse-Kemp et al., 2015). This new tool allowed the rapid construction of an intraspecific high-density F2 linkage map of the *G. hirsutum* genome ($[AD]_1$] and an interspecific high-density F2 linkage map between *G. hirsutum* and *G. barbadense* (Hulse-Kemp et al., 2015). A_2D_1 CSSL development would be facilitated if large numbers of mapped SNPs were available to discriminate between *G. hirsutum* versus A_2 and D_1 sequences, thus it would be desirable to create a similarly high density SNP linkage map from the aforementioned A_2D_1 BC1F1 mapping population, e.g., using the same SNP array. Moreover, such a map would expectedly allow for useful comparisons to other maps produced using the CottonSNP63K, e.g., to examine differences in structure and recombination rates. However, the automated genotyping based on the CottonSNP63K in previous work relied on the original cluster file that was released with the array; that file was designed for genotyping of AD-tetraploids, not tetraploids containing germplasm introgressed

from A_2, D_1 or other diploid species. Thus, an initial goal must be to develop a customized cluster file suited to the respective mapping population.

To use the CottonSNP63K SNPs cost-effectively for CSSL development, the SNPs to be used for MAS must be rendered amenable to inexpensive simplex or low-plex SNP assays. It is thus important to establish the rate as which mapped SNPs included in the CottonSNP63K can be converted to PCR-based assays.

The combination of this high-density linkage map and an efficient, simplex or low-plex genotyping pipeline will enables affordable genotyping in a targeted segment (or segments) of interest, can facilitate introgression breeding, and in the pyramiding of genes or traits of interest. Furthermore, combining the above with high-throughput DNA extraction method (Zheng et al., 2015) provides cotton breeders with a very powerful platform for introgression and downstream marker-assisted selection, e.g., for rare gene combinations, selection of very rare recombination events, or combinations thereof.

Materials and Methods

Plant materials

Two BC0F1 plants (200209097.10- 20029097.11) were obtained by cross-pollinating TM-1 (9909003.06) and the synthetic tetraploid $A_2A_2D_1D_1$ (9909002.13), hereafter symbolized simply as A_2D_1. A plant from a TM-1 line was used as the female and back crossed to the A_2D_1-BC0F1 as the pollen donor to produce BC1F1seed at the Texas A&M University campus greenhouse during 2003. In the summer of 2013, the seed were extracted from low-humidity storage and germinated in ragdolls. Germinated

seed were transferred into peat pellets (Jiffy, Canada) and allowed to grow two weeks in the greenhouse. Night temperatures were lowered for the last few days to promote "hardening", after which *ca.* three-week old seedlings were transplanted into a campus breeding nursery plot along F&B Road of College Station, Texas. Young unfurled leaves (2-4) were collected from 48 of the BC1F1 plants and 1-2 leaves per were subjected to DNA extraction using the Macherey-Nagel Plant Nucleo-spin (Pennsylvania) extraction kit following the recommended protocol by the manufacturer. Extracts were initially quantified for DNA concentration and wavelength (260/230 and 260/280, nm/nm) ratios using NanoDrop2000 Spectrophotometer (Thermo-Fisher Scientific Inc., Massachusetts) to determine the quality of each extraction. These extractions were stored in -80^{o}C freezers.

We concurrently advanced the backcross introgression breeding materials while we developed resources for linkage mapping. As the marker platform was being developed, advanced backcross progeny were developed from each BC1F1 derived line by systematically backcrossing each line in successive generations (**Figure 1**). In the summer of 2014, 20 BC2F1 lines, each derived from a different BC1F1 line, were planted in the breeding block on campus at College Station, TX, in most cases with 20 seedlings per BC2F1 family; several individual plants per family were backcrossed as seed parent with the TM-1 line as the pollen parent. Reciprocal crosses were attempted, but resulted in lower boll set, inferably due to low pollen fertility. Lint samples were also harvested for phenotypic analysis. Similarly, in the summer of 2015, seed of 42 BC3F1 families, each corresponding to a single BC1F1 plant, were germinated, and *ca.* 15

seedlings for each family were transplanted and backcrossed to TM-1 to produce BC4F1 seed. Tissue samples were collected for 8 randomly selected individual BC3F1plants per BC1F1 line were collected and stored in -80^{o}C freezer for DNA extraction and subsequent application of MAS, i.e., once the KASP platform is completed.

An initial sample set of five of these individuals were initially genotyped on the CottonSNP63K array, to determine if it might be a suitable platform for genotyping the mapping population. Once the genotyping platform was deemed suitable, it was decided that the size of the mapping population should be increased to improve accuracy of the linkage map. Therefore, additional BC1F1 remnant seed from similar backcrosses made in the year 2009 were germinated in ragdolls, transferred initially to peat pellets, and after hardening in the greenhouse, hand-transplanted to the breeding block at College Station, Texas, in the year 2015. True, young leaf tissues were sampled from 25 individuals and DNA was extracted from each with the Macherey-Nagel Plant Nucleo-spin kit following the manufacturer's protocol. It was initially quantified with the Nanodrop2000 spectrophotometer. All the 73 BC1F1 individual (**Table 2**) DNA samples were then re-quantified using PicoGreen and diluted to a uniform concentration of 50 ng/μL. The DNA samples of all 73 individuals were eventually genotyped using the Illumina CottonSNP63K array.

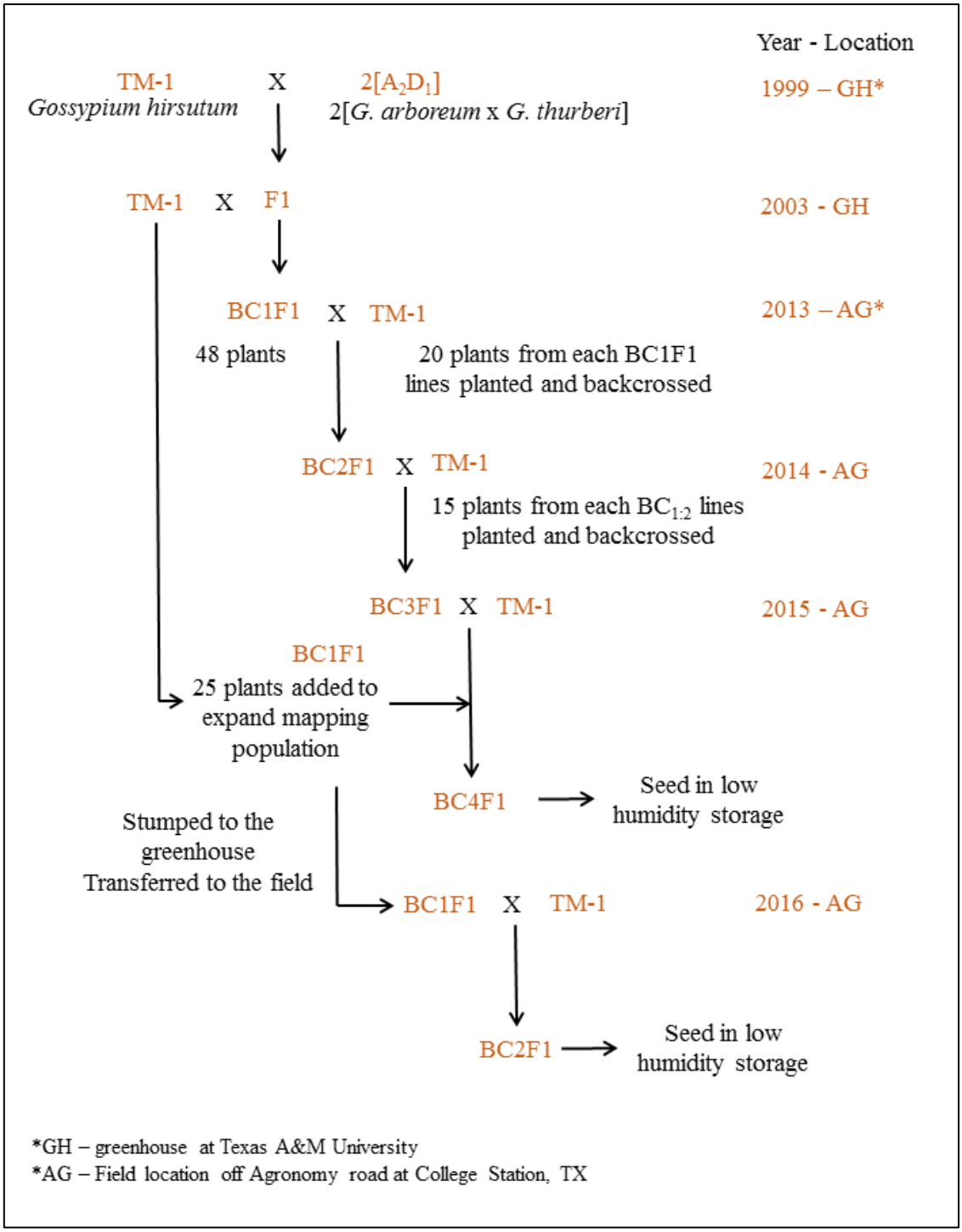

Figure 1: Breeding scheme used in generation the of A_2D_1- derived populations. The species mentioned first in each cross was used as a female.

Table 2: Pedigree information and plant identification codes of all plants included in the mapping population.

Individual ID	Female Parent ID	Male Parent ID	Female Parent Type	Male Parent Type
201208003.03	200000111.11		TM-1	
201307143.17	200600114.01		A_2D_1	
201308091.04	9909003.06	9909002.13	[[2(A_2D_1)*TM-1] BC0F1]	TM-1
201300573.14	200209099.08	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300573.15	200209099.10	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300573.18	200209100.07	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300573.20	200209100.05	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300574.04	200209099.20	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300574.08	200209099.20	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300574.09	200209099.20	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300574.10	200209100.05	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300574.11	200209100.05	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300574.12	200209100.05	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300574.13	200209100.05	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300574.15	200209100.04	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300574.16	200209098.12	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300574.18	200209098.12	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300574.20	200209099.14	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300575.03	200209099.14	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300575.04	200209099.14	200209097.10	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300575.07	200209099.14	200209097.10	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300575.09	200209098.12	200209097.10	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300575.11	200209100.06	200209097.10	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300575.13	200209100.06	200209097.10	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300575.15	200209100.06	200209097.10	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300575.17	200209100.06	200209097.10	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300576.03	200209099.20	200209097.10	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300576.05	200209099.20	200209097.10	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300576.06	200209099.19	200209097.10	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300576.07	200209099.19	200209097.10	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300576.10	200209099.19	200209097.10	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300576.11	200209099.17	200209097.10	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300576.13	200209099.17	200209097.10	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300576.14	200209099.17	200209097.10	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300576.15	200209099.17	200209097.10	TM-1	[[2(A_2D_1)*TM-1] BC0F1]

Table 2: Continued.

Individual ID	Female Parent ID	Male Parent ID	Female Parent Type	Male Parent Type
201300576.16	200209099.17	200209097.10	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300576.20	200209098.15	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300577.01	200209098.15	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300577.02	200209098.15	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300577.03	200209098.15	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300577.04	200209135.11	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300577.06	200209135.11	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300577.10	200209135.12	200209097.10	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300577.11	200209135.12	200209097.10	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300577.12	200209135.12	200209097.10	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300577.13	200209135.12	200209097.10	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300577.15	200209098.19	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300577.18	200209098.19	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300582.10	200209098.12	200209097.10	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300583.01	200209100.06	200209097.10	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201300583.02	200209099.19	200209097.10	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201500616.16	200209098.14	200209097.10	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201500616.18	200209098.14	200209097.10	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201500617.05	200209100.02	200209097.10	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201500617.07	200209099.18	200209097.10	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201500617.09	200209098.15	200209097.10	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201500617.15	200209099.02	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201500617.17	200209099.02	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201500618.03	200209099.02	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201500618.07	200209098.19	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201500618.11	200209098.19	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201500618.13	200209098.19	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201500618.15	200209098.19	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201500618.17	200209098.13	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201500618.19	200209098.13	200209097.11	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201500666.13	200908002.01	200908141.05	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201500666.15	200908002.01	200908141.05	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201500666.17	200908002.01	200908141.05	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201500666.19	200908002.05	200908141.06	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201500667.11	200908002.05	200908141.06	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201500667.17	200908002.09	200908141.07	TM-1	[[2(A_2D_1)*TM-1] BC0F1]

Table 2: Continued.

Individual ID	Female Parent ID	Male Parent ID	Female Parent Type	Male Parent Type
201500668.01	200908002.09	200908141.07	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201500668.05	200908002.08	200908141.09	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201500668.07	200908002.08	200908141.09	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201500668.09	200908002.08	200908141.09	TM-1	[[2(A_2D_1)*TM-1] BC0F1]
201500668.11	200908002.08	200908141.09	TM-1	[[2(A_2D_1)*TM-1] BC0F1]

Customizations of cluster file for automated genotyping

A "cluster file" (*.egt) is a support file that is developed to accompany Illumina BeadChip products such that computer-automated data analysis using the Illumina's GenomeStudio software can quickly generate genotype calls. This file contains numeric data that statistically describe the shape and position of each SNP genotype-specific cluster, i.e., for the homozygotes and the heterozygotes of all SNPs included in the SNP manifest. This cluster file is provided because the locations of clusters, though reproducible in most cases, vary from SNP to SNP. A standard cluster file is usually developed for every new SNP array, usually based on analysis of a diverse set of samples comprising over 100 individuals that represent the range of anticipated genotyping targets. Thus, based on a given sample, the cluster file defines expected positions of homozygotes and heterozygotes for each SNP.

For some samples, however, the SNP genotype-specific positions may fall outside the standardized cluster positions (outside the expected range), in which case automated procedure for "genotype calling" will fail. This was found to be the case

when automated genotype calling was first attempted with the A_2D_1 BC1F1 mapping population, i.e., using the Gh-GP1 cluster file that was developed based on AD-tetraploids. This indicated that positions of some SNP locus genotypes were shifted in the A_2D_1 BC1F1 mapping population and did not correspond to the respective distributions observed for the population of AD-tetraploids used by Hulse-Kemp et al. (2015). One option would be to ignore such SNPs, whereas another is to redefine the clusters for some or all of the SNPs for a given population of individuals and then re-analyze the ability of GenomeStudio® to call genotypes accurately. Were a customized cluster file to significantly improve genotype call frequency and accuracy from the CottonSNP63K, it would prospectively lead to better and denser linkage maps.

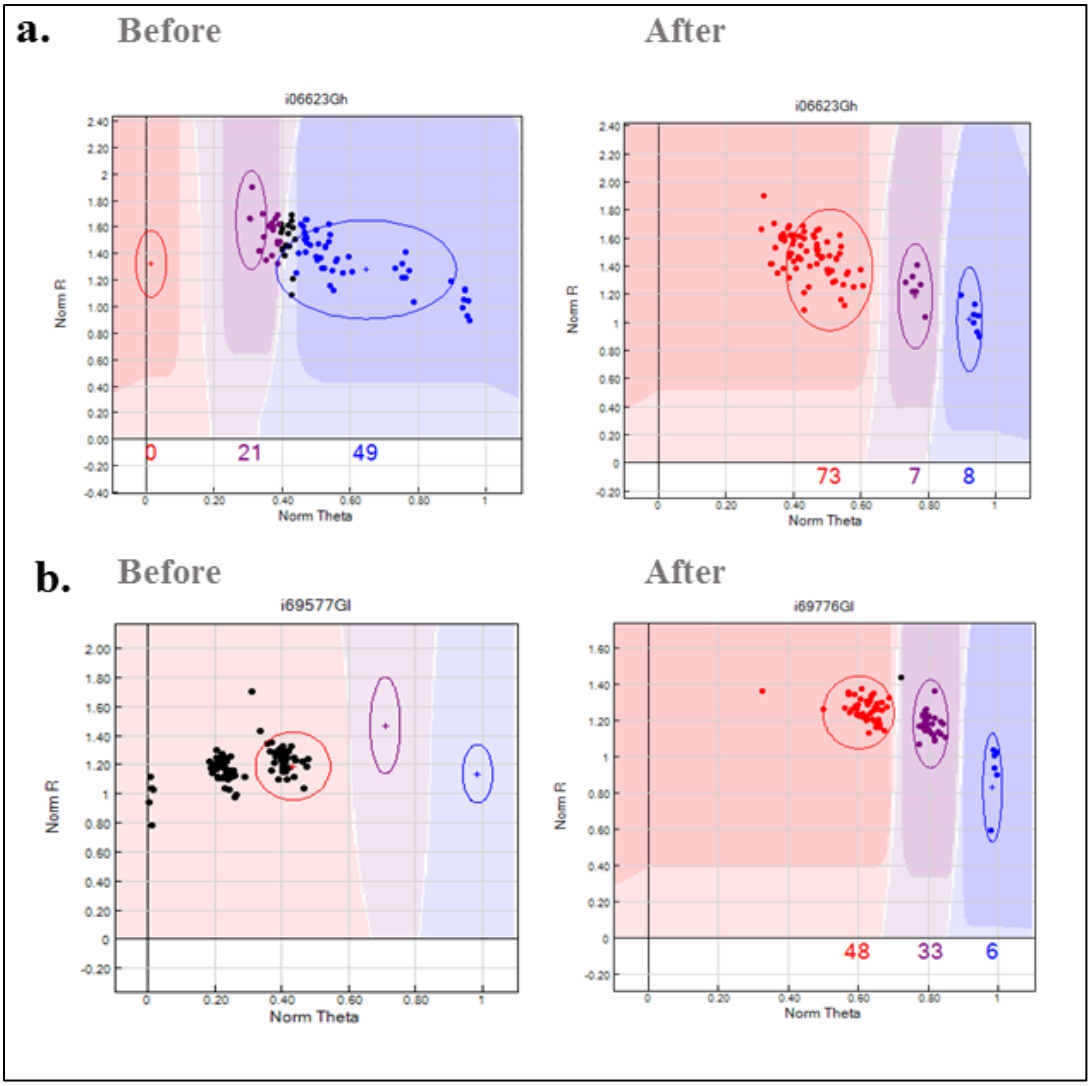

Figure 2: Effects of customized SNP cluster definition on frequency and accuracy of genotype calls. Cluster patterns observed in diploid introgressed tetraploids deviate from the expected cluster patterns defined based on AD-tetraploids by Hulse-Kemp et al. (2015), which were devoid of germplasm introgressed from diploids. (**a.**) It can be observed that this SNP the genotypes called for the samples are inaccurate and show the correct genotypes in the redefined clusters based on the sample types. (**b.**) The number of accurately called genotypes increased from 0 to 87, while 1 sample was not genotype-called.

Due to the differences in cluster positions of the diploid-introgressed individuals, the tetraploid cluster file was customized for samples of six different diploid genomes and two diploid introgression populations. This panel used is reference as the diploid-introgression diversity panel (**Table 3**). Customization of a CottonSNP63K cluster file for Upland cottons containing diploid introgressed materials was undertaken using SNP data on 69 samples involving germplasm from A, D and F genome groups (Table 3). This customization was executed following the Illumina's Infinium genotyping data analysis technical note (Illumina, 2014). The clusters for the three marker classes that can be genotyped in GenomeStudio® (AA, AB and BB) were visually evaluated for each SNP. The loci with poor performance (<1% call frequency) were eliminated from the report (zeroed SNPs). The loci with samples located in the grey zones containing over 80% of successful samples were visually evaluated to be accepted, zeroed (excluded) or manually adjusted by moving the cluster. The loci with samples in three distinct clusters that were above the lower limit (0.4) of the gray zone and not automatically genotyped by the tetraploid file were manually repositioned to improve the accuracy and frequency of genotype calls (**Figure 2**). Once all the SNPs included in the manifest were manually evaluated, this new cluster file, referenced as Gh-GP2 cluster file, was exported from the GenomeStudio® project. This cluster file was subsequently used to genotype the BC1F1 individual included in the mapping population. It is important to note that while this cluster file can be applied to similar populations, its accuracy is expected to vary, especially when the individuals involve germplasm from other diploid species.

Table 3: Diploid-introgression diversity panel: Samples included for the cluster file development.

Sample Type	Genome	Number of Samples
Inbred – *G. hirsutum* (Cultivated)	$[AD]_1$	1
Inbred – *G. barbadense* (Cultivated)	$[AD]_2$	1
Inbred – *G. mustelinum*	$[AD]_3$	1
Inbred – *G. tomentosum*	$[AD]_3$	1
Inbred – *G. arboreum*	A_2	7
Inbred – *G. thurberi*	D_1	3
Inbred – *G. raimondii*	D_5	2
Inbred – *G. trilobum*	D_8	2
Inbred – *G. armourianum*	$D_{2\text{-}1}$	1
Inbred – *G. longicalyx*	F_1	1
Synthetic tetraploid	$2[A_2D_1]$ / FADD	2
Interspecific F1		5
Interspecific backcrosses		
FADD-BC1F1		34
A_2D_1-BC1F1		5
Intraspecific F1 of *G. arboreum*		3
Total		**69**

Genotyping with the CottonSNP63K

Sample DNA concentrations determined by the PicoGreen quantification and then standardized at 50 ng/μL, then processed according to the Illumina protocols and hybridized to the CottonSNP63K array at Texas A&M University's TIGSS (Texas A&M Institute for Genomic Sciences and Society) facility. Single-base extension was performed and the arrays were scanned using the Illumina iScan. The total samples

included in the final project were spread across two individual runs performed on the same iScan instrument. The image files (.IDAT) were saved to be genotyped by the diploid cluster file. All the image files were uploaded to a single project in GenomeStudio®. The final project had a total of 88 samples, which included TM-1, A_2D_1, F1 (TM-1xA_2D_1), A_2D_1-BC1F1's, *G. arboreum* and *G. thurberi* samples.

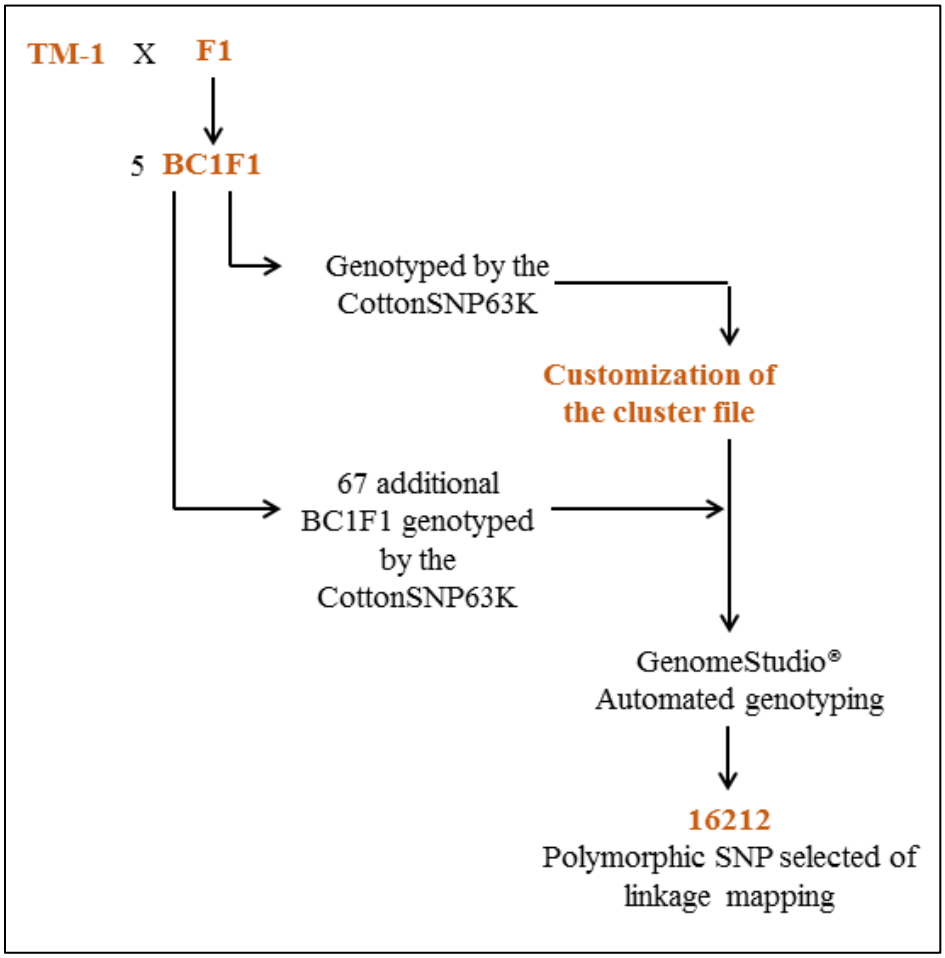

Figure 3: Workflow used in genotyping the BC1F1 plants and generation of genotyping report.

Using the GenomeStudio® software, a new "project" was created. The sample sheet, the SNP manifest and the new cluster file were uploaded to the project and a final report with genotype data was generated (**Figure 3**). A "SNP manifest" is provided for each BeadChip, providing a record of the location of each SNP on the BeadChip for all the SNPs included in the array. The "sample sheet" has the information that includes the identity of the sample, the sentrix barcode and the sentrix position of the sample; both of which are required for the GenomeStudio® to locate the image files that correspond to the sample identity, for the automated genotyping. This final genotyping report, after being subjected to appropriate quality filters (missing data, parental difference, and minor allele frequency) and other formatting steps, was uploaded to the linkage mapping software to identify linkage groups.

Genetic linkage analysis

Genotype data were transformed into map data format ("ABH") for the 73 BC1F1 individuals and the final set of polymorphic markers. The filtering steps allowed for retention of only the suitable markers with opposite homozygous allele calls between parents and therefore behaving co-dominantly, and having a call frequency of over 85%. Consequently, the data files were uploaded to JoinMap® 4.1 (Van Ooijen, 2006). After verification of segregation patterns, extremely distorted markers (p<0.0001) in the dataset were excluded from mapping. The grouping parameter of independence LOD was used for selecting linkage groups. The start LOD score was 10.0 with one point increments up to 22.0. After this value of LOD, the markers started falling into

individual groups of one each. All the linkage groups were selected at a LOD score of 12.0 or over. Once the linkage groups were selected, the map distances were estimated using the regression mapping algorithm with the Kosambi mapping function and the default parameters. The linkage groups were compared to the interspecific map of *G. hirsutum* x *G. barbadense* by Hulse-Kemp et al. (2015) and the identical markers in corresponding linkage groups were used to determine the chromosome ID of the linkage group.

Linkage disequilibrium for each linkage group was visualized using the CheckMatrix 2D plot analysis (http://www.atgc.org/XLinkage/). Discordant linkage groups (**Figure 4**) were re-estimated using the maximum likelihood mapping function and if the map order produced acceptable 2D plots, the map order was used to re-estimate the map distances with the regression mapping function. If the linkage groups remained discordant, they correlated with the interspecific map order of *G. hirsutum* x *G. barbadense* and it was observed that the regions of discordance corresponded to the anomalies observed in map orders. Therefore the framework of the interspecific map order of these markers was used as the fixed start order and the re-estimation of map distance for these groups under the regression mapping function was performed. Their accuracy was re-verified using the linkage disequilibrium plots. The map order with the most acceptable 2D plots was considered to be the final map order. The number of recombination events per individual and the average number of recombination bins across the linkage map were also calculated, along with the numbers of SNPs per bin.

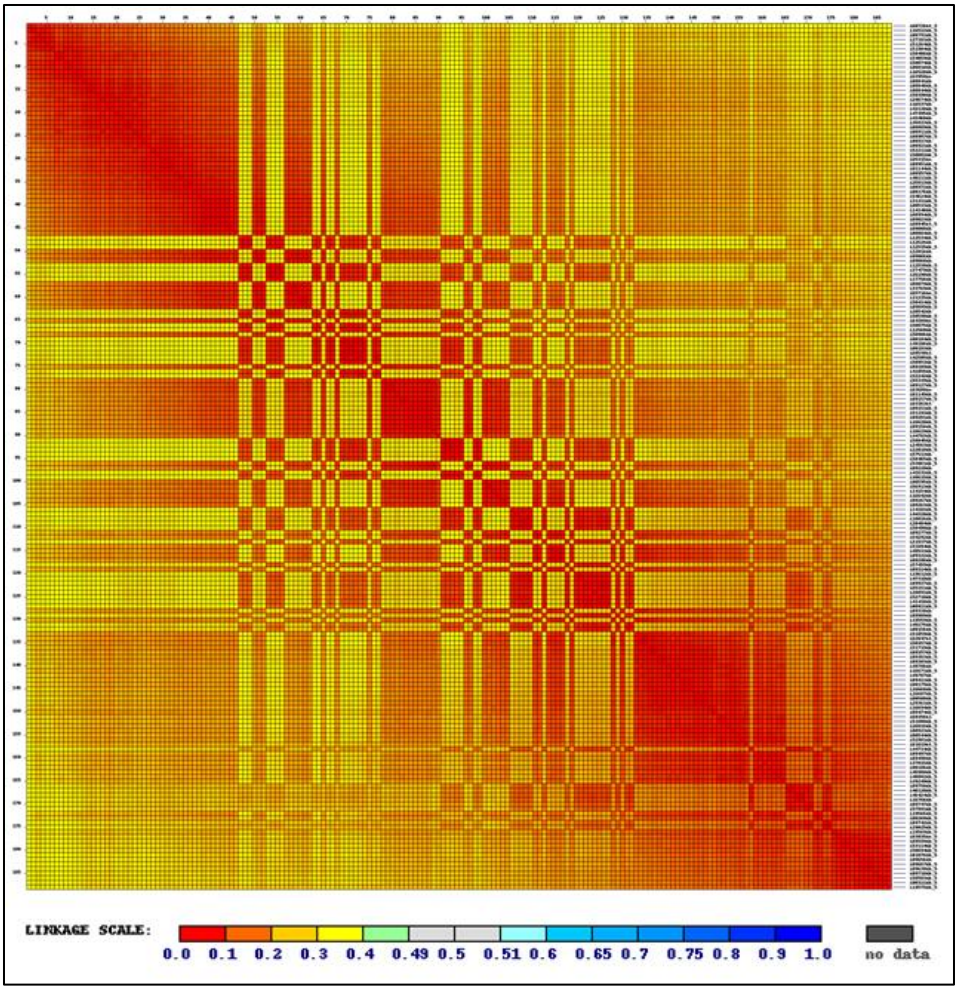

Figure 4: Example of a discordant linkage group as observed during the preliminary analysis for Chr05. Comparison with the published interspecific map indicated that difference in relative map order could be responsible for the anomalous 2D plot.

Introgression analysis

From the final map of 14,411 SNP markers developed for the A2D1 - BC1F1 population, a framework map of 1,969 markers was derived. When compared to the

complete map, the framework map offered a regular coverage of all linkage groups at a density of one marker per cM.

Introgression analysis of the BC1F1 population was performed using the CSSL Finder software version 0.9.722 (http://mapdisto.free.fr/CSSLFinder/). The default parameters were edited to suit the project (**Table 4**). Given that the population was the BC1F1 generation, the desired segment size was indicated as 50 cM (minimum,10 cM; maximum, 90 cM) while permitting large overlaps and using the minimum tiling path to select a subset of lines providing optimal coverage of donor genome in the recurrent background. It was decided to not allow single locus (marker) segments. Heterozygous regions were considered as donor regions. The percentage of wild genome in the BC1F1 generation, the mean size, number of introgression segments per line and per chromosome were also estimated.

Table 4: Settings for introgression analysis in CSSL finder.

Parameter	Setting	Description
Population & Markers		
Lines	75*	Number of lines in the data file
Markers	1969	Total number of markers in the data file
Chromosomes	26	Number of linkage groups in the data file
CSSLs Search		
Minimum segment size	10	Minimum desired size of introgressed segments
Desired segment size	50	Desired size of introgressed segments
Maximum segment size	90	Maximum desired size of introgressed segments
Large overlap	Yes	Permit overlaps of more than one marker between segments
Minimum tiling path	Yes	Look for minimum number of segments of desired size to cover the genome

Table 4: Continued.

Parameter	Setting	Description
Allow one-locus segments	No	Force to choose one-locus only fragments in case that no fragment was found at the minimum size
Two-pass search	Yes	Optimize CSSL searching
Treat heterozygotes as donor	Yes	Consider heterozygotes as homozygotes to search segments. Also alters ANOVA computation.
Complete missing data*		
Double recombinant threshold	0.001	Double recombinants of probability > threshold will not be completed
Mapping function	K	H: Haldane K: Kosambi
Megabase/cM ratio	0.224	Used when distances are expressed as megabases
*75 lines include 2 replicates. Genotype is annotated as A for TM-1 (*G. hirsutum*) allele, B for A_2D_1 (Synthetic tetraploid), H for heterozygote and "-" for missing data.		

Simplex SNP assay validation panel

A random set of 52 SNP markers (**Table 7**) that were mapped in the linkage map were selected at the rate of 2 markers per chromosome. Primers were designed for KASP SNP assays (LGC Genomics) using the BatchPrimer3 software (parameters for primer synthesis - primer type: allele specific and allele flanking primers; Tm: minimum, 55^o; optimum, 57^o; maximum, 60^o; max difference, 2^o; product size: minimum, 20 bp; optimum, 25 bp, maximum, 30 bp). Primers were synthesized by IDT (Coralville, Iowa, USA) and working aliquots of the assay mix were diluted were according to the KASP developer instructions (LGC Genomics, Hoddesdon, UK). These primers were run on the "A_2D_1- SNP validation panel" (**Table 6**) which contained 24 samples (**Table 5**), including TM-1 (Stelly Lab) (x2), 2[A_2D_1] (Stelly Lab) (x2), F1 - TM-1x 2[A_2D_1] (x2),

16 different BC_1F_1 individuals and water non-template controls (x2). The plates were subjected to PCR in either a thermocycler (GenePro, Bio-Rad Laboratories Inc, Hercules, CA) or a hydrocycler (LGC), depending on the availability. The PCR program was as follows: hold at a temperature of 94^{o}C for 15 minutes to denature the DNA; followed by 10 cycles of 94^{o}C for 20 s and temperature of 57^{o}C - 65^{o}C for 1 min. After these cycles, additional rounds of 28 cycles of 94^{o}C for 20 s followed by temperature of 57^{o}C for 1 min. Plates were read by Pherastar (BMG Labtech, Ortenberg, Germany) to read the fluorescence intensity at default wavelength after 38, 44 and 50 cycles. These files were later analyzed using the KlusterCaller (LGC) program. The SNPs were labelled as "co-dominant" if the assays produced three clean clusters that allowed for differentiation and scoring of the parents and the F1 genotypes in individual clusters. The SNPs were labeled "dominant" if there are only two clear clusters and the F1 genotype was indistinguishable from one of the parental clusters. The SNPs were labeled as "failed" if the PCR assay resulted in no amplification or if no definable clusters were observed. (The definitions of "co-dominant", "dominant" and "failed" are used throughout this thesis.) The "dominant" and "failed" markers were classified as non-functional markers that would not facilitate subsequent marker assisted selection efforts. Those markers with unclear clustering patter were re-run to confirm their classification.

Table 5: Sample types and their identities used in the SNP validation panel.

Annotation	Sample Type	Sample ID	Sample Conc. used
P1-GH	TM-1	201308001.04	10 ng/µL
P2-A_2D1	A_2D_1	201307143.16	10 ng/µL
F1	F1	201308091.06	10 ng/µL
Sample-1	BC1F1	201300573.14	10 ng/µL
Sample-2	BC1F1	201300573.15	10 ng/µL
Sample-3	BC1F1	201300574.12	10 ng/µL
Sample-4	BC1F1	201300574.04	10 ng/µL
Sample-5	BC1F1	201300574.15	10 ng/µL
Sample-6	BC1F1	201500144.01	10 ng/µL
Sample-7	BC1F1	201500145.10	10 ng/µL
Sample-8	BC1F1	201500146.02	10 ng/µL
Sample-9	BC1F1	201500144.04	10 ng/µL
Sample-10	BC1F1	201500146.03	10 ng/µL
Sample-11	BC1F1	201500144.07	10 ng/µL
Sample-12	BC1F1	201500146.07	10 ng/µL
Sample-13	BC1F1	201500144.08	10 ng/µL
Sample-14	BC1F1	201500144.10	10 ng/µL
Sample-15	BC1F1	201500146.10	10 ng/µL
Sample-16	BC1F1	201500147.01	10 ng/µL

Table 6: SNP validation panel plate map.

	1	2	3	4	5	6	7	8	9	10	11	12	
A	NTC	P1-GH	P2-A_2D_1	F1	Sample-1	Sample 2	Sample-3	Sample-4	Sample-5	Sample-6	Sample-7	Sample-8	**SNP-1**
B	NTC	P1-GH	P2-A_2D_1	F1	Sample-9	Sample-10	Sample-11	Sample-12	Sample-13	Sample-14	Sample-15	Sample-16	
C	NTC	P1-GH	P2-A_2D_1	F1	Sample-1	Sample 2	Sample-3	Sample-4	Sample-5	Sample-6	Sample-7	Sample-8	**SNP-2**
D	NTC	P1-GH	P2-A_2D_1	F1	Sample-9	Sample-10	Sample-11	Sample-12	Sample-13	Sample-14	Sample-15	Sample-16	
E	NTC	P1-GH	P2-A_2D_1	F1	Sample-1	Sample 2	Sample-3	Sample-4	Sample-5	Sample-6	Sample-7	Sample-8	**SNP-3**
F	NTC	P1-GH	P2-A_2D_1	F1	Sample-9	Sample-10	Sample-11	Sample-12	Sample-13	Sample-14	Sample-15	Sample-16	
G	NTC	P1-GH	P2-A_2D_1	F1	Sample-1	Sample 2	Sample-3	Sample-4	Sample-5	Sample-6	Sample-7	Sample-8	**SNP-4**
H	NTC	P1-GH	P2-A_2D_1	F1	Sample-9	Sample-10	Sample-11	Sample-12	Sample-13	Sample-14	Sample-15	Sample-16	

Table 7: Map positions and SNP sequences of markers selected for KASP primer design and assay validation.

SNP ID	Chr	A_2D_1 Map_Pos	Allele A	Allele B	Inter_GhxGb_ Map_Pos	SNP Sequence
i52306Gb	AD01	43.697	T	C	23.1	ATTTGTTGTCTTCAAAATTTCTTCATTAGAAGTAGCATAAAG AAGTGGATYTTGATGTTTTCTTATGTAGATATGGATTGGTTT AAGTGGGAGTTAGCTCT
i30614Gh	AD01	87.434	T	G	103.64	TTCATTAGAATTATTCATTCAATATACTCTCTCAAATCTTTGT TTTTCATKCTTCTTCATCAAATCAAATTTCTGCTTGTTTGTTT GCCAAGTCTCAAGCC
i03066Gh	AD02	0	T	G	0	AAACTTGATATTTGCTGTGCAGTAAACTTTTCTTTTTAATGTT AACAAGTATTTCTAATTTACTGCCAGACCAAGCGCATTGGAG GTGTTGCGTCATCCTYTATTTTGGAGTTGTGAGATGAAACTG TCTTTTCTTCAAGAGACTAGTGATAGGGTTCAATTAGAAGAT AGGAAGGTTGACTCTGACATCTTGAAAGCATT
i29065Gh	AD02	53.317	A	G	73.53	ATAACACCCAAACACCCGTAAATGATCATATGTAGGGTCAC TACCATGAARAACTTAATACGGAGATTGACCTTTCAAAACCA AGGTAGGTGGACGATTAA
i23426Gh	AD03	7.244	T	C	N/A	TTAGCCTCTGTTACTATATAGTTCCCCAAACATCTACCAAAT CTCATGATYTATCCTATCTAACTGTTACACTAAGCCTCATATT GCAGCAAATTCTTTCT
i54909Gb	AD03	69.73	T	C	93.38	GTTTGCCTTTTGGCAGCCGTCRATGATTGATGACTAACCTCC ACTTCCCTYCTATTCCTTACAACCAAAGTGTTAGTCTTTTTTT TGGCTAGGACATCTTC
i43499Gh	AD04	37.299	A	G	N/A	ATAGTGCATGTATCTTTGTCAGCCATTTTCCATTAATAAAGC TTATAAGCRCTCTTCCAAAATTTCCTAGTATTTCTTGGGAGTA AGAGGAAAGATTGAAG
i24385Gh	AD04	80.39	T	G	0.85	GATTCGCTTCGTTTTGGTTTGTTACCAGTTTGYCACTCATATT TTCATACAATATCCTAACCTCAGCCTAAGTACCCACGGTT
i09277Gh	AD05	120.035	A	G	159.91	TCCAAAGCAACTCTGAAGTTTTGGCTTCAGAGAATTCCTTTG ATCCGTACAGAGTACGTGGAAAATTAAGCTGGGCTGAAATC GGAAATTATGGATTATCRACTGAAGTATCTTGGATGTCAGTT GGGAAGCAGCAGTTGGAATATGCATCTGGGGCCTTAAGGAA GTTCAGGTATTAAAGCTTGTCAAACCTTTTTGATT

Table 7: Continued.

SNP ID	Chr	A_2D_1 Map_Pos	Allele A	Allele B	Inter_GhxGb_ Map_Pos	SNP Sequence
i09151Gh	AD05	131.367	A	G	173.15	GCGTACCCCAGAACTCCATGGATGATTACTACGAAAGAAAA AAGGAAGGGGCTAGGGTTTTGCGAGGAGGAGATAAGTAGG GTTTTAGAAGGAGAGATACRACAGAGTAAAGCTCTTTCCGA AGAGGCGGAGTTGAGTGGGGGCGTTGAAATGAATTTATAAG CAAATTGCCGCCCTCTATTCTATTTTTAGGGTTTCTTT
i32075Gh	AD06	110.694	A	G	122.52	GAAAAATGCAAAGTTCGTTCGTTGGTCGTTGTAGAGGTGCTA CTACTTTGRTCACTGTTGTTGTTGTTGTATCTAGGGAGAC ATTCAACCAACGTTTC
i61821Gt	AD06	111.309	T	G	N/A	TTAAATCTCCGAGGAACTTGAACCATGTTGAAAGCAAAGGT GCTTGGTTTKTCTATGGCTTTTTATGTCGACTTTCTTGTTTTTT GTTCTTATATTCAATG
i18226Gh	AD07	72.334	A	C	79.16	TTGATACGGGGACGGACTTGTCAACTTGGCGTTCGATTTCAT ATACTTTTTAAACTATATTCGCGCTATAAACTTAAATGCATT AAAATTAAATATGGACMAGTGTGTTCCTTTGGAAGAGATTA TGAATTCATAGGACTTAGGATAATATTAAAAAAAAAAACAT GATTGTGTCTGTATATAACTAGTTTATATGAGCAA
i41382Gh	AD07	105.34	A	G	45.46	TGTTGGTTTCGAGTTTTCGGTTTAATCGTATTTTAAGGCTGAT TTTGAGCRTTGGTATGTTGACTCTTAGGTTCTAGAGGGCTCC TGGTTGCTGTAGAATC
i49570Gh	AD08	0.074	T	G	2.99	ATGATGGTTTATTNGCTGATTGAAGGTGGAGGACAAACTAG GTTAATTAACACATTAAGAGCTGGCTGTATAATCTGCAAGTA CACAATAAAATGGAGGTGAATGTTAGTAATTAAACCAGTAA AACTACAAATCAATATCTGCTTTCAGYGTCTGAAAATAGCAA CTATACTTCCAAAGCCANTAACACAATATCACATAAAGGGN NCTTCAAGTAAACANCTAGCATTATTTGCTANGAACAAGAA TAACAATAAGAACAACATTACAATATCACAGCATATGCGCA CCTCAGGGTCCA
i30796Gh	AD08	30.458	T	G	34.26	CAAACTTTTTTTACTGGATCGAAAAACTGTGCAAGGTTAATT GAAGTAGAKAGGACCTAAACAAAGTTGGGAAGCAATTGAAC ATAGTAGGTATTAAGGAT

Table 7: Continued.

SNP ID	Chr	A_2D_1 Map_Pos	Allele A	Allele B	Inter_GhxGb_ Map_Pos	SNP Sequence
i49356Gh	AD09	4.253	T	G	23.93	GACTCCTTAATTGCTTGCCAATTGTTTAGAGATCACTTAATC CTTGACTCYTTTAGGTTCAATCCTTGGAATACTTGGGTGTTCC ATTAAACATTACGAAT
i59188Gb	AD09	9.054	T	C	33.31	CTAAAATTGATCGATTAGTCGAAAACTTTGTCTTTCCCTCGA TTTAGATCYGAGTTTCACTGTTCTTGATATAAATGTAATCAA AATTAACTTATTTAATC
i29528Gh	AD10	79.673	T	G	81.69	CTTCACGAAATGGGTAGAGACATCTTCTTATACTAGTGTCAC CAAGTCAGYAGTAGGGCGATTTTTGAAAAAGGAGATCATTT GTCGGTATG
i00854Gh	AD10	90.467	A	G	103.49	TAGTGTAAAATAAGAAAACACATCAATAATTTTATTCATCAC AAAACAAAAGTTTGACAGAATGCACTACCCTAATTAATGAA TGAATGAATGAATAAACMCCCCTAAAAAATGACTACTCTCC AAATTCGTCCTTTTTAACTTATGAAAAGAAATTGATTCATAC CATCAACTTGTAGATCTAGATTTCAAATCCCACCT
i07000Gh	AD11	41.929	T	G	41.84	AACGGTCTTCCCCGCAGTCTCTGATCCTATGCATTCGAAATG CCGTACTCCTTTGATCTCCACCGTCCATTTCAAACGACAGTT ATCTTCTTCACCGACAYTCCCTCGAGAAACAACCCAATTACG CGCGCTGAAGCTGGAACCCAATCCATCTCTCTCGTTGAGTAC ACTCGGAAACACGGTGTCGGCGCTGCATCCCGG
i65080Gm	AD11	150.408	T	C	178.93	TAATAGAACACATGGAAACCAAATGATATATATATATGGTG AGGCCTCTCYTCTACTTCCCTGCTAATTTCTGATATATATAAT ATATATACAATACTAGC
i48341Gh	AD12	55.751	A	G	50.6	AGAGCTACGATTTAGTCGGTTCTAGGACTGTCGTAGAGATGT TAAAGTTCRCTATACATGCCATTATTTGAATCTTAATAGTGT GACGTCTCCTAATTGTT
i52587Gb	AD12	59.894	A	G	65.62	TAGAAATCGGTGTCGGAAGCGGTGAGTTCCGACAAACAGGA CAAGACCCATTAAGTTTCARCCAAGCATCAATACAAGTTACA TGAAAACAATGCCTACATTCAGGCATCATCCTCAACAT
i61586Gt	AD13	47.582	A	G	75.35	CAACAATCACAACAAAGAGTCATGCTCTAGGTTAGTGATTC GTAACAGTAMAGTAGAGGTTATCTGTTGTAAAACGAAAATG AACTTGAACATACCGTACT

Table 7: Continued.

SNP ID	Chr	A_2D_1 Map_Pos	Allele A	Allele B	Inter_GhxGb_ Map_Pos	SNP Sequence
i15353Gh	AD14	114.389	T	G	138.11	ACTTGTAAGAACTCCATGTTGATTGGTTATTTGATGCGGGTA AGATTTCTACCATATTGACATGATTTGAATGGAAAAGGTTAC AAAGAAGCATAAAACAYTTCAAGGAGAGTTAAATTAAGCCA TTCTTGCAAGGTTGAGACCTCGATTAGCTTGCTTTTTCCAATT CTGACTCATTGAGTGGTGCATTATTGGTATGAA
i02792Gh	AD15	37.705	T	G	42.82	ATTTTGATTCTTCAGAAATGCCTACAAAGCTCTATTCTTCAAT CTTCCCATCTATCTTCTTTCTTTCCCTCATTTTCCTCTTCCTTT ACTCCACTTTCCTYCCTCTTTATACCAATAACGACTCCTCTTC ATTGCCCACTAATAAGTTCCTTCCTTCATCCTCTCCTCCTTGT AACCTTTTTAAGGGCCATTGGGTTTTAA
i18484Gh	AD15	40.574	A	G	42.82	ACTTTGGCAAGCTTTCAGCCCTTGAATCCTTATCAATGAGGG GGGTCCAATGGTGTTGGGATGCAATAAGCAAAATGCTAGAG TGGTCTAGTGAGGTGAARCATCTCTACATGAAGGTTGAATTC ACTGGAGATTTGGAGTCCCTTTTACCCTTTCCAGAAGTCGAT TTCGTTGAGTTTTTTAACAGCCATCCCAAGCTGC
i01980Gh	AD16	54.2	T	G	41.2	TGCTCCTTTTCTGCATTATCAGGTTTTGGAGGAGTACAAGCC AAAGATATAATCCTCAAACAAGGGAGAATTATATGCTCCGA TATGGCAGGATGCTTTGYACCCAATTTTATGGAAGAAAATA ACAGATGGAAAACAACACGCAATCTTGATTCCCAGAATTCA TCAGCCAGGGAACACACTTCTGAAAGCAATAACAAT
i00934Gh	AD16	98.034	A	G	103.58	TTCAACAGGATCACCATCTATTGACACATCCATAAAAACTAG AGGATTCTTCTTCTTTGCCATCCTTGCTCACCAGATTCACAAC TCAAGCCAGTACAACMAACAATTTCGTCAAAGCATTGCCAA ATTCACCAATCTTCATTAAAGAATATAGCAAAGCAGAAAGG ACCTTTTCCCCTATATATGCATTATTTCTTTCGT
i37731Gh	AD17	40.455	A	G	45.75	AAGCCTTTTAGATACCTAATCAAAACACTAATCAACAACATG GAAGTCAAMTAGCCACTTGTGTGTGTATTGAATTCTTAATTG CATAATTTTGATTAATT
i39720Gh	AD17	40.462	A	G	45.75	AAATTGGAAAGATTCTGGAATAAGTAGCTCAGATTATCATC ATCTTTTTARGTTCCCTATCAAAGATACCAAACCCGAGTAGA AGGCATTGTAAAACCTAG

Table 7: Continued.

SNP ID	Chr	A_2D_1 Map_Pos	Allele A	Allele B	Inter_GhxGb_ Map_Pos	SNP Sequence
i32363Gh	AD18	50.903	A	G	54.48	AATAGGGTGTTAGAATAAGAGAAAAGAGGAGAAAAAGAAA AGAAAAAAAAMGAGAGAGAGCAAATTCTATTAGGACTTGG ATAGGGGAAAAGTGCAAAGTC
i63450Gm	AD18	73.331	T	G	N/A	GCAGCTTGGGCATTAGCCAACCTCTCAGCCGGACAATTTGTT ATGGGGGCYGAGGATAATGTGAGTTCAAGATTGCTTCCTGA TATTTCGGTTAATTCAAC
i09418Gh	AD19	84.659	A	G	N/A	TTCAATACAGGTTTAATGAAGCTCCCGTTTGATCTCCCCGGT TTCGCGTTCCGAAACGCCAGGCTGGCCGTTGAGCGATTAGTC GAAACCCTTAGTGATTRTGCTACTCAAAGTAAAAAGAGGAT GTCTGAAGGAGACGAGCCTTCTTGTTTAATCAATTTTTGGAT GCAAGAAACTGTTAGAGAAATAGCGGAGTCCAAA
i23637Gh	AD19	101.194	A	G	199.28	ACCTGAGCCACGGCGAATTAGGTTATTTATATATGGGTTTAG ATTTCGCTMGGCCCATTACTAATTGGAACCATATTACCCGGT AAGCCTTTTAATGGGCT
i56485Gb	AD20	47.217	T	G	N/A	GTAAAAACTTTAATATACTTTTAACAAAATAGTTGTCGTGAC ATCACTATYGAACTATTCTAGATGGCAGAAATGAGAGTTTAT AAGAGTCGAATGGTGGT
i12092Gh	AD20	49.475	T	G	48.79	TTGCAAGAAGGTCTTCTAGTATCACTGGTGGAATGGGCAAA GAAAGAGATGGATCATCTAAATGGGAAGTCAGTCCAATAGA ATTAATGTTGAGAGATGCKGAGCCTTATACAACTAGCGTTCG AATTGGCAAGGGTTTTCAAGCAGAAGTTCCTGACTGGTCTGG TCCAATTGATATGTATGTTCCATGAACTAATTATT
i51853Gb	AD21	70.001	A	G	103.92	GTTTCAGGTGATACCGATACCTTGTAAGTTCCTGGCGAGCCC ACATTGGTRAGAGTTCGAGTGTGTTTAACAACATTCGACCCA CCCATGATCGAATCAAA
i07420Gh	AD21	84.355	T	G	131.78	TTTAGGTTATTAATTCTTACACAACCCACAGAAGAAATTAAA ACAAAGTCTTGTCAACATTTTTCATATAGATTTCTTGTCTCTT AATAATTTCTTCTCGKTTTTTTCTTCATGCTGATGAACACATT CAAACTCCCTGCTCGATCCCTTTCGAACCGGTTTGCTGCTCTC TTTTCTTTTTATCCTAGTAACACTTGATTC

Table 7: Continued.

SNP ID	Chr	A_2D_1 Map_Pos	Allele A	Allele B	Inter_GhxGb_ Map_Pos	SNP Sequence
i51144Gb	AD22	3.022	A	G	26.17	GGTCATTCTTTCATGGGGAATAGCATCAGAAGCCTCCTCAAA CAGGTTGTRTAGCACACTCATACAGGATTGACAAACAGCAT AGAAATCCTCCTTGTTGA
i12867Gh	AD22	83.673	A	G	N/A	GACATCTTGACTATATTCGAACTGTGCAGTTTCATCATGAGA ATCCTTGGATTGTGAGTGCCAGTGATGATCAGACTATCCGCA TATGGAACTGGCAGTCRCGAACTTGTATCTCTGTGTTAACTG GTCATAATCATTATGTTATGTGTGCATCATTCCATCCTAAAG AGGACCTTGTCGTGTCGGCCTCCCTTGATCAGA
i05827Gh	AD23	9.853	T	G	4.7	ACCCTTGGATCAGAAGAAGATGATGATGGGTTCTCTGGGTTC TCTTTAAGGGTTACAAGGAAGATCTTCCTTAGGATCATATCT TCTATTTCTTCTGGGGYTCTTTGGGGTTTTTGCGTAGCCATAG TTTCGAAGACAAAAAAGGGAATAAAAACTTGGTTTTTTCCCT CCTTTATTTTCCTGAGAAACAAATAGGGTTCT
i34362Gh	AD23	88.681	A	G	108.28	AATTGTTCTCGCATTAAGGCTTTAACTTTTCATTGTCCAAGTT AGTTTCTRATATTGGTAATTGTTCCCATGTTAAGGCCTAAGC TTGACGATTGTTCCCA
i04583Gh	AD24	41.215	A	G	31.68	TTCTATGGAACAAAACGGTGGACCAGATATTTGTCCCAGTAA TTGAAGCCATTAGTTCCCGTCGATGGTAGAAGATTTGGGGTA ATTTTTACTGTCAACTRTTGGTGTGGCTCTTACTTCAGATCAT TGCATGCGTACAATTGCTTGCTGTTAATTTTATCTCTCAAATC TCTTTGAACCAAAAGGGGGAATGCTTACTTG
i38741Gh	AD24	84.677	T	G	76.96	TTTGAATTCTTTCCACAATTAAGATCTCTTGAGAACAACGAA TAACAACTKAAATAAAATCTAAAACCTAAGAACAAAAGAAA GACTAAACTAAATTTACT
i10927Gh	AD25	80.259	A	G	69.76	TGCGGTCCTAGGCTAAAGGGAAGTTGACAAAGGCCAATAAA GCTTTAGCTAAAGTGCAGGAGCATCTTGATCCCACAGATCTC CCGACTGATTTGGAAACRTTGAGTGAAGAGGAGAGAATTTT ATTCCGTAAGATTGGTCTGAGTATGAAACCCTACTTGCTTTT GGGTAAGAATGGGAACTTCAATCCTGCTTTACTGT

Table 7: Continued.

SNP ID	Chr	A_2D_1 Map_Pos	Allele A	Allele B	Inter_GhxGb_ Map_Pos	SNP Sequence
i17070Gh	AD25	123.517	T	G	130.85	TGTTGGTAATCGCCACGAAGTTTGCCTTGCGACGGAGGGTAA GATACCTCCATAGCACCGAAAAACGTAGCCGTAGAAACTTT GATTTGGTTCCCTAAGCYTGCTTTAACTATTGCTTTATGAAC GTTCTGCATTGCTGGGAAAAGGAACCGAGCAAATGGATCCT GTGGCTTTACTTCGTTCCCAATAGTAGCGACGTAT
i56086Gb	AD26	58.761	A	G	60.18	AAATGATAGCAAGAAGAGTAAGAGAATGTTTCAAGGACCCA TTGTAAAAARAAGAGAAATTGATGGTGAATTTTAACGGTTTA GGCGGCAGCACACAAGGG
i47939Gh	AD26	133.342	T	G	145.65	CATTGATTATCGATTTCAAAGGTAAATAAAAAAACACCACC CTTCCATGCKCATTTCTTGTTTCATTTTATGTAGGTTAATTTT GACGAGTTAAACTTGAA
i38406Gh	AD13	85.402	T	C	107.37	AGCTCGTGATTTTGTTAGAACATCTAGTGATCGAGATGATTT TGAATCTTYTGTTAATGTTAGTGATTATGTGGCTCAACTTCTT TACAAGAAACCCCTCA
i56884Gb	AD14	67.466	T	G	71.61	ATAAGGTCATTGATGTGCTTCATCCAAGATTTTTAGCTTTAA CTTTTCTTYGTTGGGAACATATAGCCTTCCCATGAAGCGTAA TTCTCCTTCTTTACCCA

Results

Automated genotype calling with diploid cluster file

All six distinct clustering patterns described for functional markers characterized across the AD-germplasm (Hulse-Kemp et al., 2015), were observed across the diploid-introgression diversity panel. In Pattern-I clustering (**Figure 5a**), all samples fall in to a single cluster. The probe sequence for these markers detects a monomorphic locus or loci. The Pattern-II clustering (**Figure 5b**) detects two clusters that are detecting two monomorphic loci, in which each cluster is homozygous for a different allele. These markers are hypothesized to correspond to intergenomic SNPs or "Homeo-SNPs", and therefore appear to be heterozygous in all lines. Functionally polymorphic SNPs constitute the remaining patterns. The markers that show three clearly defined clusters and behave like classic co-dominant markers with three possible genotypes (AA, AB, BB) are classified as Pattern-III clustering (**Figure 5c**). The homozygous clusters are each located near 0 and 1 on the X-axis in the SNP graph. Pattern-IV clustering (**Figure 5d**) also involves three clearly defined clusters, but they are offset to one side of the graph with one of the homozygous clusters corresponding to 0.5 on the X-axis of the SNP graph. This pattern is hypothesized to detect two loci, one polymorphic and the other monomorphic, likely in homeologous chromosomes, where the three clusters correspond to the three genotypes (AAAA, AAAB, AABB). Pattern-V clustering (**Figure 5e**) involves three definable clusters that are quite close together. Pattern-V markers most likely detect several loci, of which only one is polymorphic and the others

are monomorphic in the tested sample. The final Pattern-VI (**Figure 5f**) clustering includes markers that have extremely close clusters and are often set as failed. The progression in complexity from types III to VI (e.g., Figure 4c to 4e) inferably reflects increasing numbers of amplifiable loci due to polyploidy, paleopolyploidy, localized duplications and(or) possibly sequence conservation among more distantly related loci. Markers that exhibited Pattern-IV- or V-type clustering (e.g. **Figure 5d or 5e**) required manual readjustment of cluster definitions to enable accurate computer-automated genotyping by GenomeStudio.

Based on the individuals diploid-introgression diversity panel, it was observed that these lines showed deviations from the cluster positions defined by the tetraploid cluster file. A customized cluster file was therefore developed as part of this project for research on tetraploid cottons containing germplasm from diploids. For convenience, this new cluster file will be referred to as the "Gh-GP2 cluster file". The A_2D_1 BC1F1 mapping population was genotyped using the both the "Gh-GP1 cluster file", originally reported by Hulse-Kemp et al. (2015) and the "Gh-GP2 cluster file" (reported here).

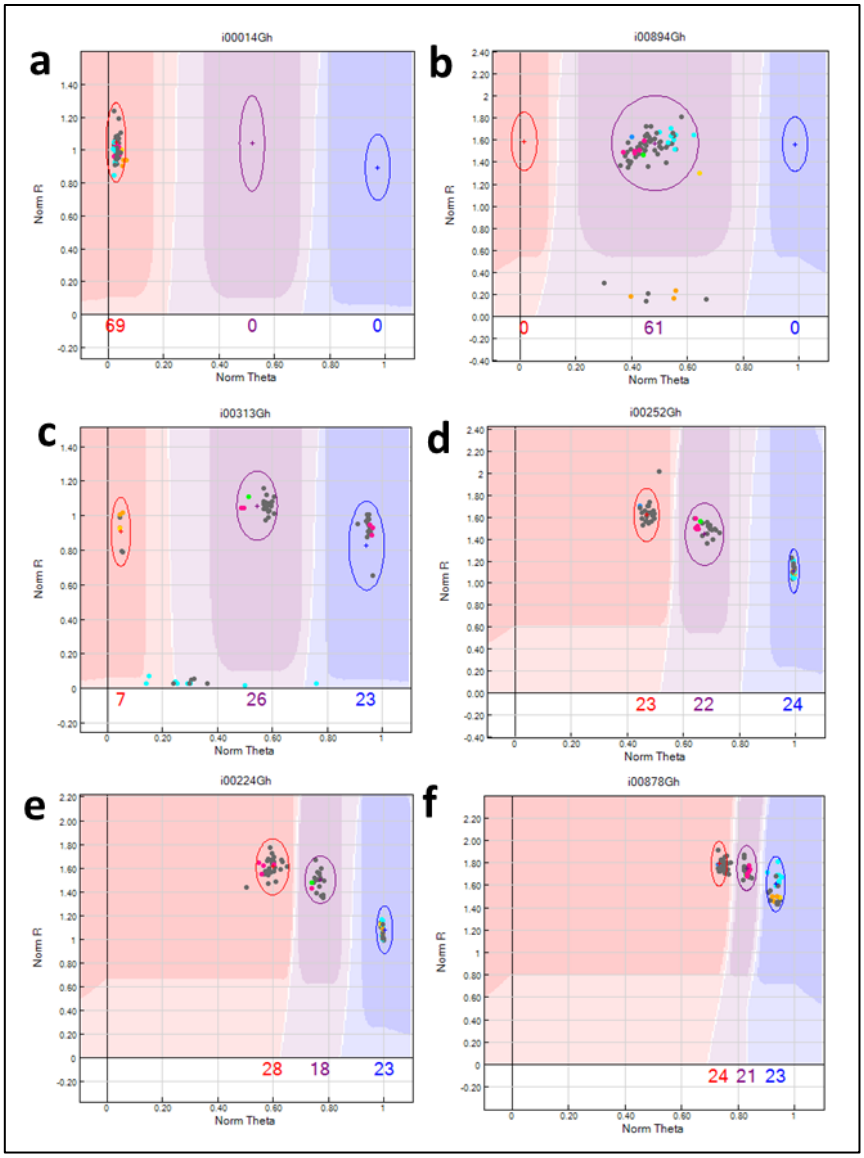

Figure 5: Classification of scorable SNP markers as published before by Hulse-Kemp et al., 2015. (a.) Monomorphic markers; Pattern-I (b.) Inter-genomic or homeo-SNP markers; Pattern-II (c-f) polymorphic markers. (c.) Pattern-III; the homozygous clusters correspond to 0 and 1 on the X-axis of the SNP graph (d.) Pattern-IV; the cluster positions are unequally distributed to one side of the graph with one of the homozygous clusters corresponding to 0.5 on the X-axis of the SNP graph. (e.) Pattern-V; the cluster positions are unequally distributed to one side of the graph. The clusters are closer than in pattern IV, but still distinguishable. (f.) Pattern-VI; the cluster positions are also unequally distributed to one side of the graph, but very close and often indistinguishable. Hence, they are often considered as failed.

Comparisons revealed that the call frequency and the total number of markers deemed polymorphic increased with the newly developed "Gh-GP2 cluster file", over the previously released Gh-GP1 cluster file. Comparison were made for the locus summary reports generated for the 63,058 SNPs included in CottonSNP63K array for the A_2D_1–BC1F1 mapping population generated using the Gh-GP1 cluster file to the report generated with the Gh-GP2 cluster file. It was revealed that the number of zeroed SNPs (excluded from automated genotyping either due to non-amplification or relative difference in the cluster position) changed from 7,920 SNPs when the tetraploid cluster file was used versus 5,384 SNPs when genotyped with the customized cluster file. This indicates an increase of 2,536 SNP markers (4%) where some or all of the samples could be called to a genotype. It was also noted that the SNPs assigned a call frequency of 1 (where all samples included in the sample sheet could be assigned to a genotype) increased from 25,495 SNPs when genotyped with the tetraploid cluster file to 28,757 SNPs when genotyped with the customized cluster file (**Figure 6**).

After exporting the final genotyping reports with both the cluster files, they were subjected to the same standard quality filters to generate the final data set suitable for linkage mapping. It was observed that the final data set included only 13,432 polymorphic markers when genotyped with the tetraploid cluster file, whereas the customized file could classify 16,212 markers to be suitable for linkage mapping. This increased number (2,780) of discernibly polymorphic SNP loci indicated a 20% improvement in numbers of callable polymorphic loci for the A_2D_1- BC1F1 population.

Additionally, after the completion of the linkage map, it was observed that 2,197 of these 2,780 loci could be mapped, which corresponds to 15.5% of the total mapped markers.

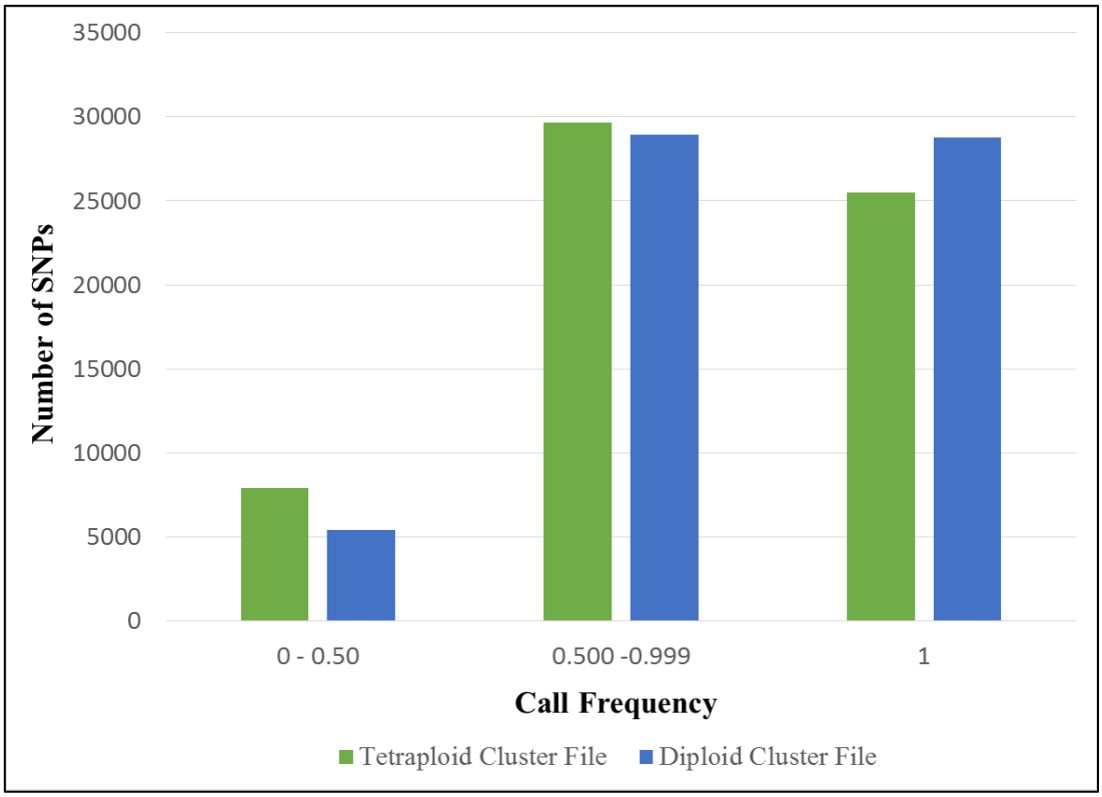

Figure 6: Comparison of the distribution of call frequencies of all the SNP markers included on the CottonSNP63K array, genotyped with the tetraploid and the customized cluster files.

The final genotyping report from the diploid cluster file initially identified a total of 19,781 SNP markers. These were classified as functionally polymorphic within A_2D_1-BC1F1 mapping population and were subjected to additional quality filters (MAF - Minor Allele Frequency and spurious genotype calls) to determine a final set of 16,212

SNP markers that was subjected to linkage analysis. A total of 14,411 markers were mapped to 26 linkage groups that correspond to 26 allotetraploid chromosomes.

Segregation distortion has been previously reported in cotton (Lacape et al., 2003; Mei et al., 2004; Shen et al., 2007). The high levels of segregation distortion in interspecific crosses can be hypothesized to be attributed to the divergence of the species (Tanksley, 1988). In map developed from the BC1F1 population of an interspecific cross of cultivars of *G. hirsutum* and *G. barbadense*, which are thought to have diverged less than 1-2 MYA, it was observed that 8% of the markers used showed a segregation distortion and that the 82% of the distorted loci mapped to c6, c12, c15, c12 and c20, where c = chromosome (linkage group) (Lacape et al., 2003). It was expected that the allele transmission would favor the elimination of donor alleles, however, the transmission of the *G. barbadense* was favored at these distorted loci. In an independent study involving a recombinant inbred lines of a intraspecific cross of *G. hirsutum* (7235xTM-1), significant levels of distortion were observed at 52.5% of the marker loci examined (Shen et al., 2007). These high levels of distortion were attributed to the introgression of *G. anomalum* (B_1) into the cultivar 7235 (Shen et al., 2007). These distorted markers were mapped to A-subgenome chromosomes A3, A10, A11, and A12, (i.e., chromosomes 3, 10, 11 and 12) and to the D-subgenome chromosomes D6, and D8 (i.e., chromosomes 24 and 25) (translation of reported LG names in **Table 8**). The high-density interspecific map of *G. hirsutum* and *G. barbadense* showed significant segregation distortion of the 18.1% of the markers mapped (Hulse-Kemp et al., 2015). Although it was reported that male (*G. hirsutum*) was favored by a ratio of 1.13:1 to the

female (*G. barbadense*) parental allele, no significant overall bias toward one of the parents was reported.

Table 8: Annotations of A- and D- subgenome chromosomes and their corresponding allotetraploid chromosomes (K. Wang et al., 2006).

A-subgenome Chromosome	Corresponding allotetraploid chromosome	D-subgenome Chromosome	Corresponding allotetraploid chromosome
A1	Chr01	D1	Chr15
A2	Chr02	D2	Chr14
A3	Chr03	D3	Chr17
A4	Chr04	D4	Chr22
A5	Chr05	D5	Chr19
A6	Chr06	D6	Chr25
A7	Chr07	D7	Chr16
A8	Chr08	D8	Chr24
A9	Chr09	D9	Chr23
A10	Chr10	D10	Chr20
A11	Chr11	D11	Chr21
A12	Chr12	D12	Chr26
A13	Chr13	D13	Chr18

The A_2D_1-BC1F1 mapping population is derived from a wide cross, so segregation distortion was expected. Indeed, significant deviations from the expected were observed at some loci. The average of the observed locus genotype frequency across the selected polymorphic loci used for mapping was 1 for the female allele (TM-1) to the male allele (A_2D_1). An initial set of 462 markers was found to be extremely distorted ($p<0.001$) and so all affected markers were eliminated from the data set used for linkage mapping. Of the remaining 14,411 markers that could be mapped, 3,207 (22.2%) showed significant distortion ($p<0.01$), and 1,951 (60%) of those underwent a

decrease in the number of heterozygotes, with a favored ratio of 1.6 of the paternal allele (A_2D_1) to the maternal allele (TM-1). This indicates that the recurrent parent allele was recovered a higher rate than would be expected if recovery were random ($p<0.01$). An increase in the number of heterozygotes with a favored ratio of 1.7 of the female allele (A_2D_1) to the male allele (TM-1) was observed for 1,256 (40%) of the remaining markers. This indicates that the donor alleles at these loci were recovered at a higher-than-random frequency ($p<0.01$). Chr10, Chr23, Chr14, Chr02, Chr21, Chr01, Chr12, Chr03 and Chr18 of the A_2D_1 linkage map had over 30% of their mapped loci under segregation distortion and accounted for over 65% of the distorted loci. Of these, homeologous relationships exist only between Chr02, 03 and 14, where 02 and 03 are segmentally homeologous with 14. Although the percentage (22.2%) of distortion in segregation observed is high in the A_2D_1-BC1F1 population, it is within the range of segregation distortion of 8% to 52.5% previously observed in the markers in other interspecific crosses (Lacape et al., 2003; Shen et al., 2007).

Genetic map construction

The interspecific map was generated from 72 individuals of the BC1F1 generation of cross with *G. hirsutum* line TM-1 to the A_2D_1 synthetic tetraploid. A total of 14,411 markers were mapped to 26 linkage groups. Linkage disequilibrium and recombination plots were generated using the CheckMatrix 2D plot analysis (http://www.atgc.org/XLinkage/), as means to facilitate the detection of incorrect SNP locus orders. Whereas the order of loci in most (20) linkage groups seemed to be

internally congruent, incongruities were observed in maps of six chromosome: Chr05, Chr12, Chr15, Chr16, Chr20 and Chr24. Correctional analysis using the framework map order from the interspecific map of *G. hirsutum* x *G. barbadense* (Hulse-Kemp et al., 2015) was performed by using the order as the start order in JoinMap® to rectify these discordant map orders of these chromosomes.

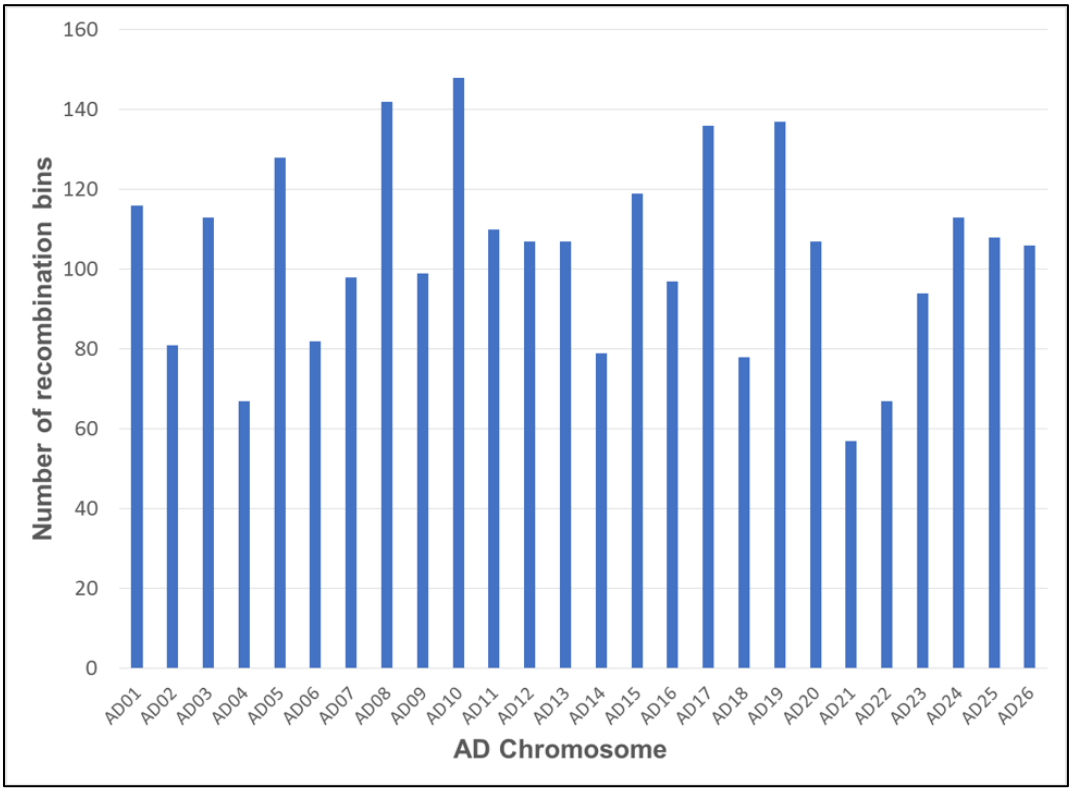

Figure 7: Distribution of number of recombination bins across the chromosomes.

The final map length, including the reordered linkage groups with 14,411 markers mapped to 26 linkage groups, was at 3,170 cM. A larger portion of the mapped SNP markers, 9,722 markers (67.4%), were from the *G. hirsutum* set. The remaining 4689 markers (32.6%) were from the other species (2,941 *G. barbadense*, 644 *G. mustelinum*, 622 *G. tomentosum*, 350 *G. armourianum* and 132 *G. longicalyx*). The average marker density 4.54 markers per cM and the largest gap observed was 14 cM. An average of 527 markers per linkage group was mapped to the A-subgenome chromosomes and 581 for the D- subgenome chromosomes. The map contained an average number of 104 recombination bins with an average density of 5.4 markers/bin per chromosome across the linkage groups (**Figure 7**). Linkage groups of A-subgenome chromosomes averaged 107 recombination bins per chromosome and a density of 4.8 SNP markers per bin, whereas linkage groups while the D-subgenome averaged 100 recombination bins apiece and had a density of 5.9 SNP markers per bin. The average length of A-subgenome linkage group maps (130.9 cM) was about 16% larger than the D-subgenome average (112.9 cM). It was observed that the 2,721 (18.9%) of mapped markers were also mapped in the intraspecific map of F2 mapping population of *G. hirsutum*. 10,902 (75.6%) of the mapped markers were also mapped in the interspecific map of *G. hirsutum* x *G. barbadense*. 3,212 markers (22.2%) of the total mapped markers are unique to this map. The linkage map were drawn using MapChart© (Voorrips, 2002). The linkage maps are referenced as AD01 through AD26 in the figure (**Figure 8**) or Chr01 through Chr26 in the text.

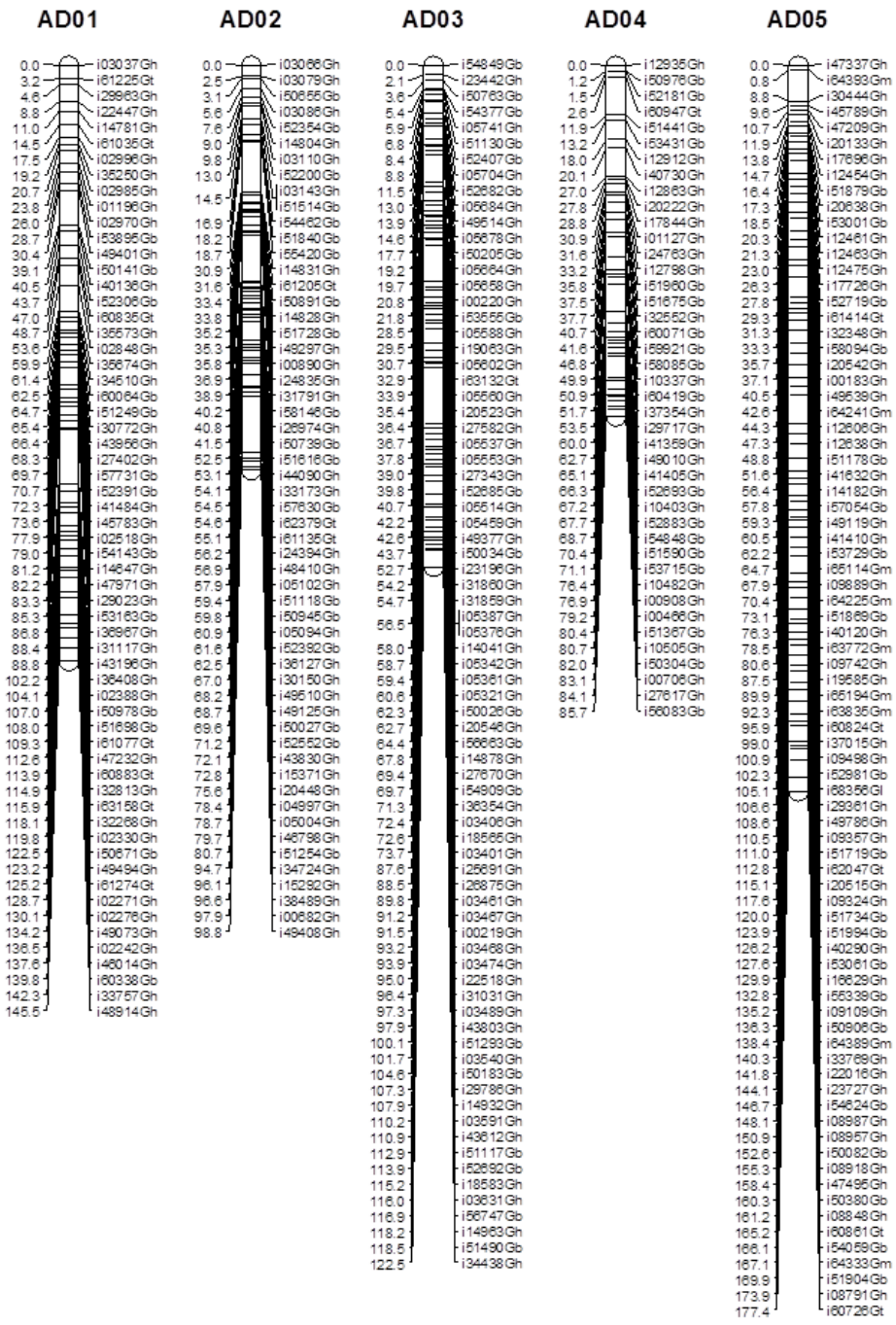

Figure 8: Interspecific linkage map of 26 chromosomes. Map determined using 73 BC1F1 individuals from a cross between TM-1 and a F1 derived from TM-1 crosses to A_2D_1 Synthetic tetraploid. No more than one marker is listed on the right per centiMorgan (cM) on the left. Chromosome numbers are listed based on AD chromosome number.

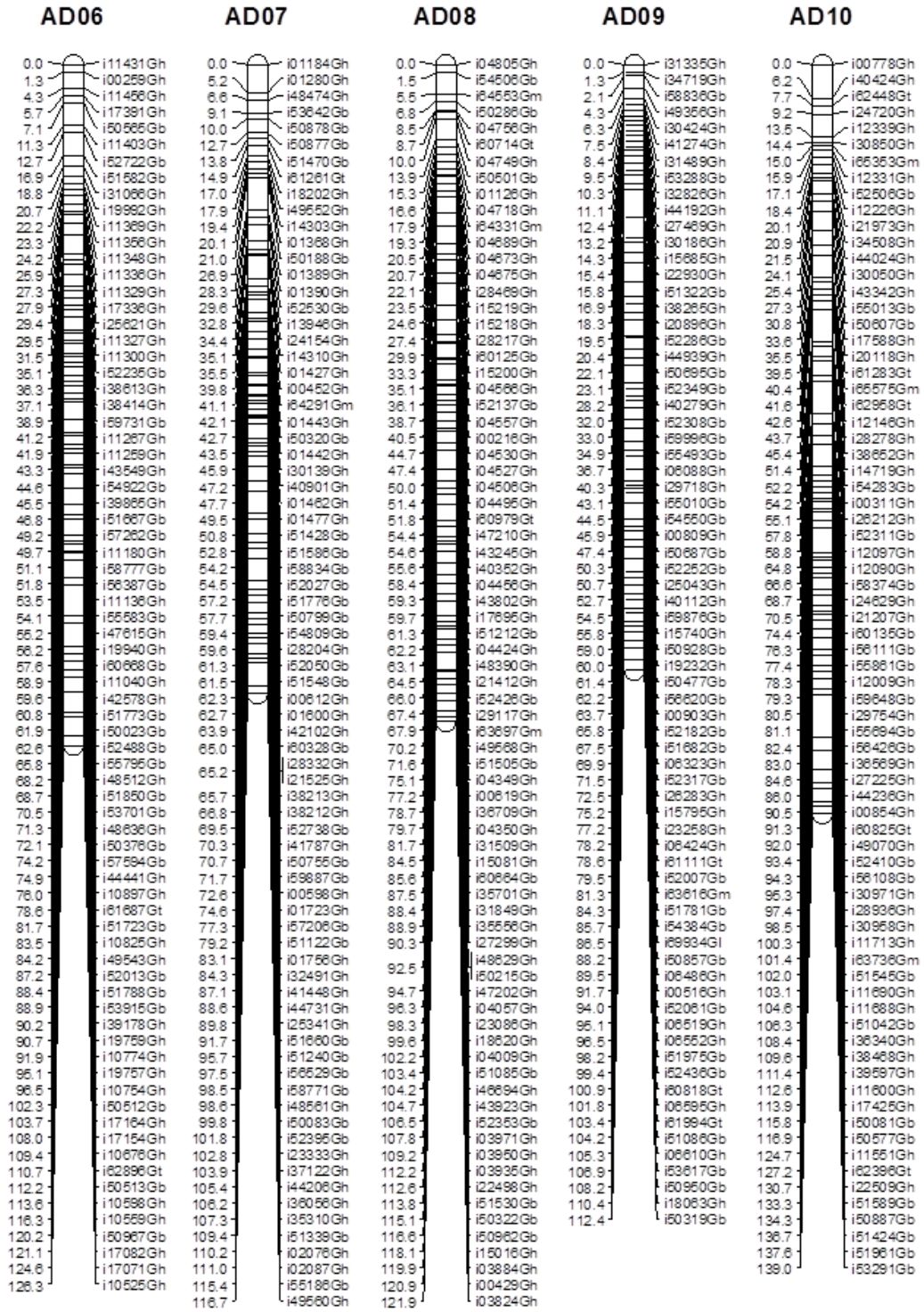

Figure 8: Continued.

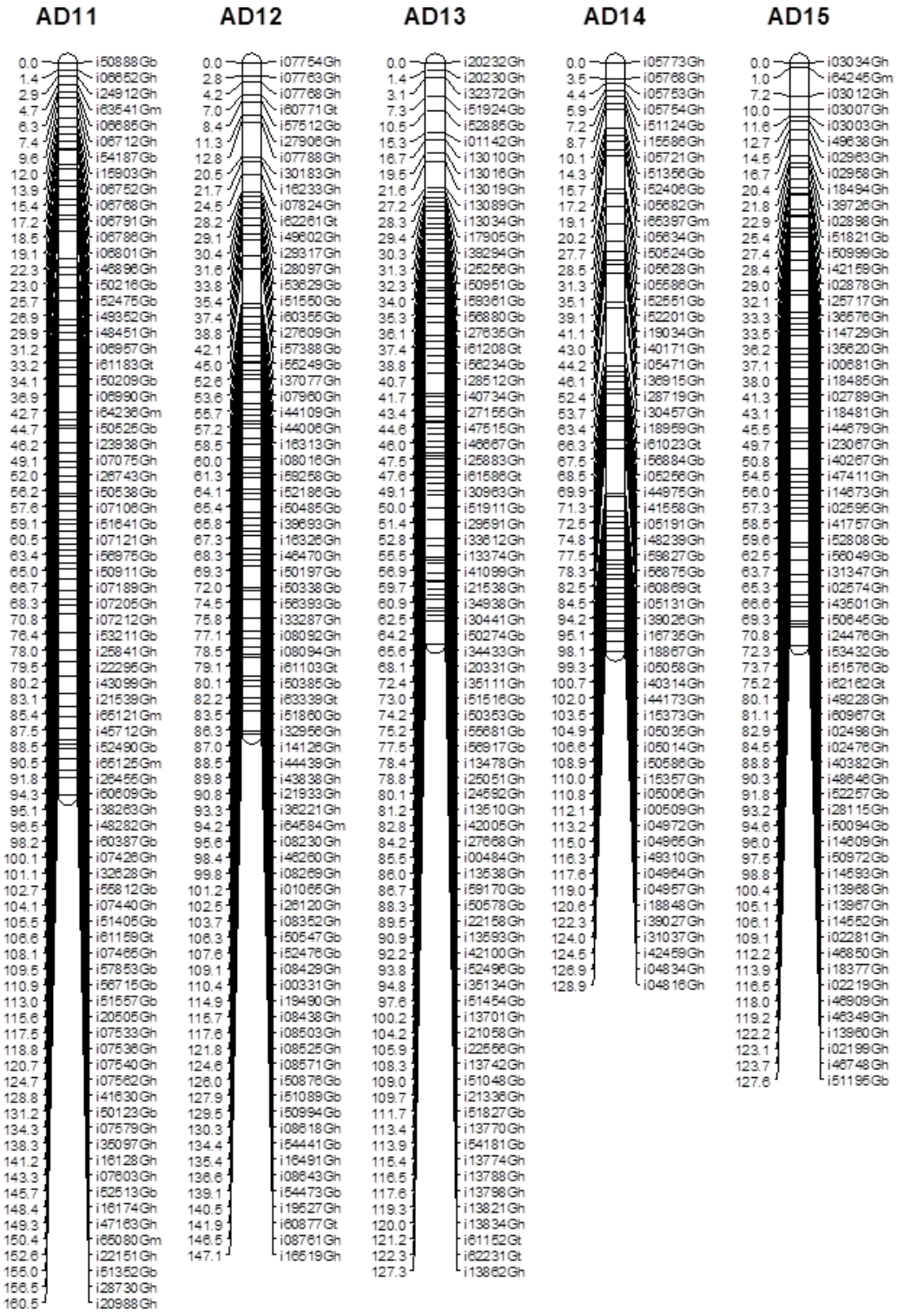

Figure 8: Continued.

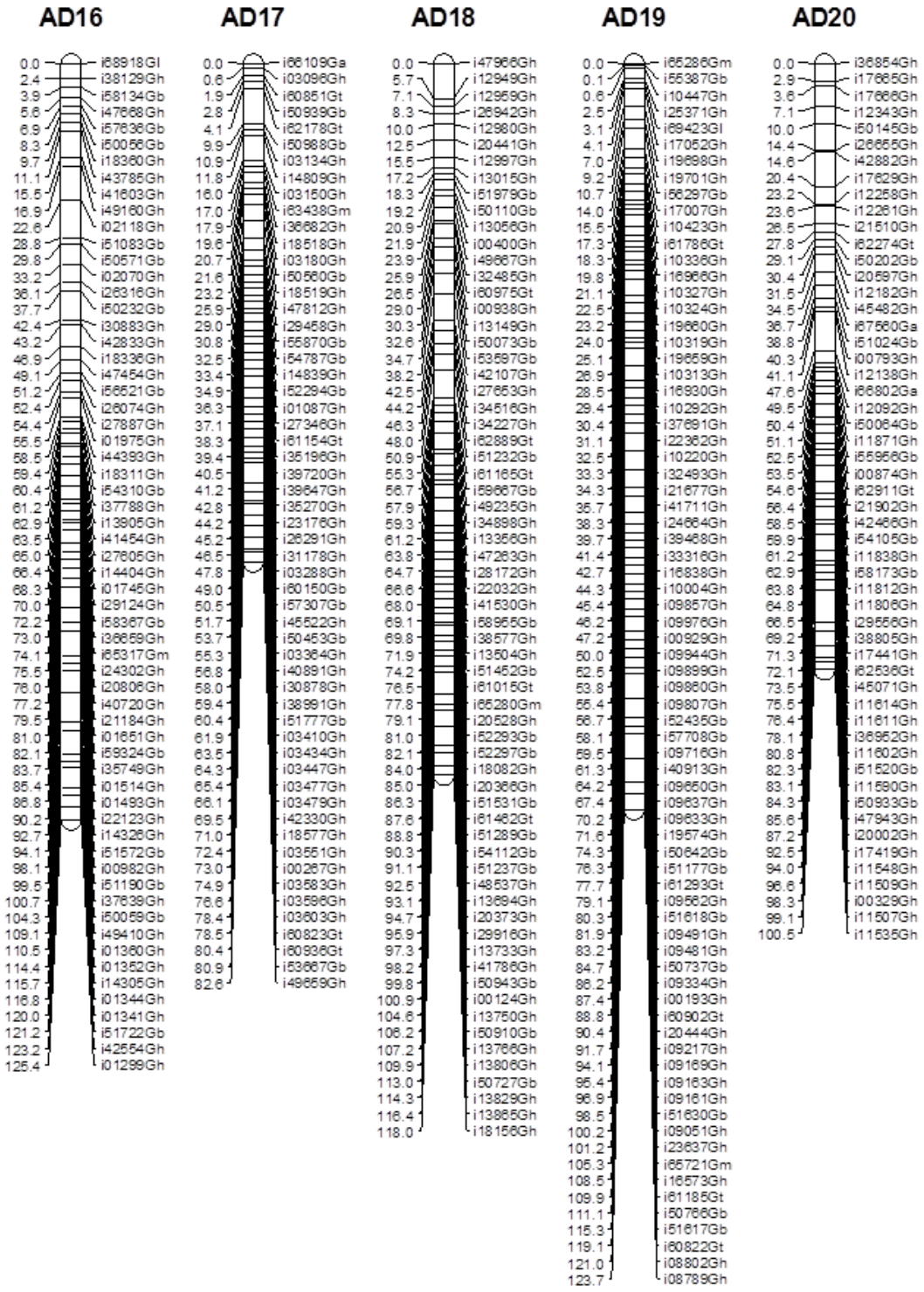

Figure 8: Continued.

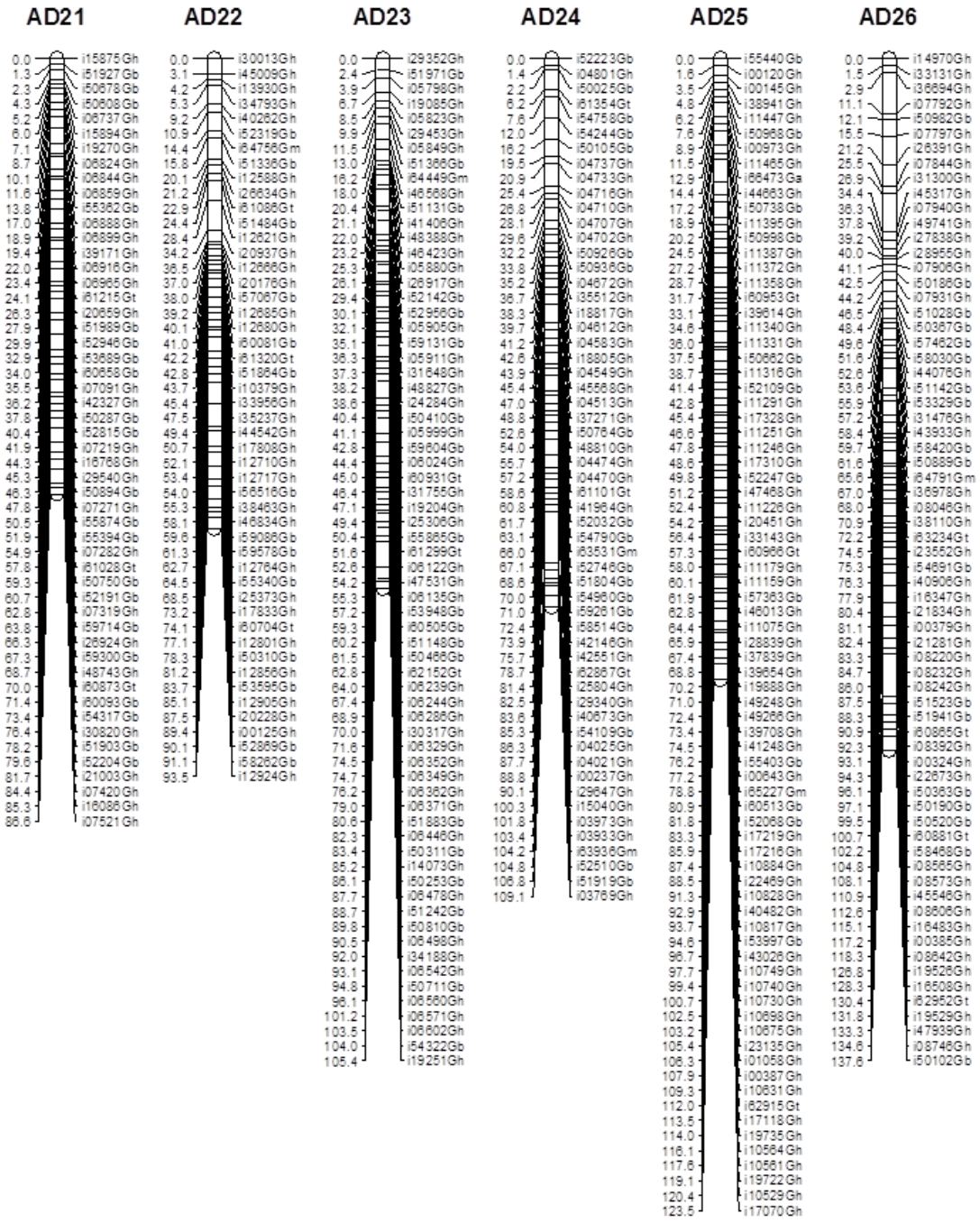

Figure 8: Continued.

Introgression analysis

In the BC1F1 generation (**Table 9**), the percentage of heterozygous donor (A_2D_1) genome per plant varied from 34.6% to 64.5% (an average of 48.7%). This percentage is only slightly inferior to the expected 50%. Of course, the recurrent parent (TM-1) genome percentage correspondingly varied from 35.5% to 65% (an average of 50.02%), and this is in accordance with the expected 50% of the recurrent parent in a traditional backcross. The distribution of lengths of the wild segments in BC1F1 for the 26 linkage groups was calculated and the lengths of introgressed segments varied from a minimum of 2.32 cM to a maximum of 89.86 cM.

Table 9: Individual statistics of genotype composition of the BC1F1 lines.

Line #	% donor	% recurrent	# of segments	# of chr w/segment
1	47.4	51.2	30	22
2	47.4	51.2	30	22
3	46.7	52.2	37	25
4	40.0	59.2	33	23
5	47.3	51.0	50	23
6	44.0	54.6	44	22
7	46.2	52.8	30	20
8	49.8	48.8	44	25
9	57.0	42.3	31	25
10	45.9	53.3	31	22
11	51.3	48.6	29	25
12	55.0	43.4	41	25
13	62.2	37.3	37	24
14	52.8	46.5	34	22
15	49.2	50.0	33	21
16	58.8	40.9	33	24
17	49.6	49.2	39	25
18	48.4	51.4	36	23

Table 9: Continued.

Line #	% donor	% recurrent	# of segments	# of chr w/segment
19	54.7	44.2	33	24
20	49.4	49.9	31	24
21	41.8	57.9	28	20
22	54.7	44.6	32	24
23	45.5	53.9	27	22
24	36.1	62.5	38	23
25	50.7	47.7	33	24
26	48.8	50.6	36	24
27	46.2	52.9	32	23
28	39.8	59.2	31	20
29	50.1	48.5	28	24
30	48.2	51.4	29	22
31	50.4	49.2	27	24
32	51.6	48.1	36	23
33	49.8	49.5	38	25
34	64.5	35.5	31	24
35	47.2	52.4	25	20
36	42.5	56.8	21	19
37	44.5	53.8	36	22
38	54.4	40.9	52	26
39	37.5	59.7	38	23
40	35.9	61.3	31	21
41	48.9	49.8	31	22
42	47.1	51.1	38	23
43	41.1	56.8	33	21
44	42.3	56.6	30	22
45	53.0	43.9	40	25
46	53.0	45.0	40	25
47	51.7	47.1	38	24
48	50.3	48.5	26	23
49	47.2	51.2	42	24
50	48.0	50.3	38	24
51	40.6	58.0	27	22
52	37.9	59.9	36	21
53	48.6	49.3	56	25
54	45.4	53.0	38	24

Table 9: Continued.

Line #	% donor	% recurrent	# of segments	# of chr w/segment
55	51.9	47.0	32	25
56	49.5	49.5	28	24
57	47.4	52.2	29	22
58	49.3	50.1	30	24
59	53.0	46.0	44	25
60	53.7	46.3	27	23
61	39.8	57.8	36	20
62	48.7	50.9	28	21
63	37.6	59.6	28	21
64	47.6	51.1	31	22
65	48.7	49.9	34	24
66	50.0	48.5	30	24
67	52.4	43.4	56	25
68	50.6	47.8	36	25
69	48.7	48.9	39	25
70	53.3	44.5	44	24
71	45.2	51.3	43	24
72	49.0	47.3	52	26
73	58.5	39.9	48	26
74	53.8	46.0	31	24
75	34.7	65.0	25	22
Minimum	34.7	35.5	21	19
Maximum	64.5	65	56	26
Average:	**48.7**	**50.11**	**33**	**24**

The average length of introgressed segments was 50.11 cM across all the lines (**Figure 9**). The number of wild introgressed segments per BC1F1 line varied from 21 to 56, with an average of 33. The number of chromosomes with one or more wild introgressed segments varied from 19 to all 26 chromosomes, with an average of 24 chromosomes having introgressed segments (**Figure 10**). The amount of wild introgressed segments in the recurrent parent background was considered to be optimal.

The number of A2 and D1 segments that were not represented in the population was zero.

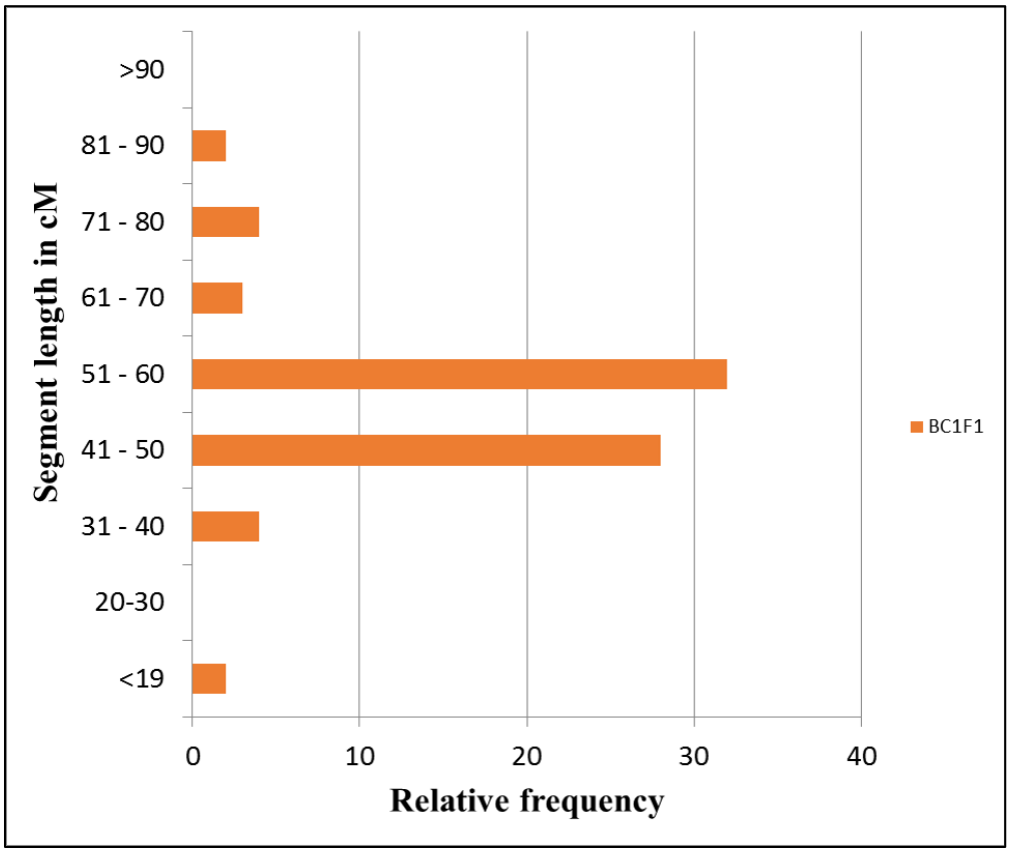

Figure 9: Distribution of introgressed segment size.

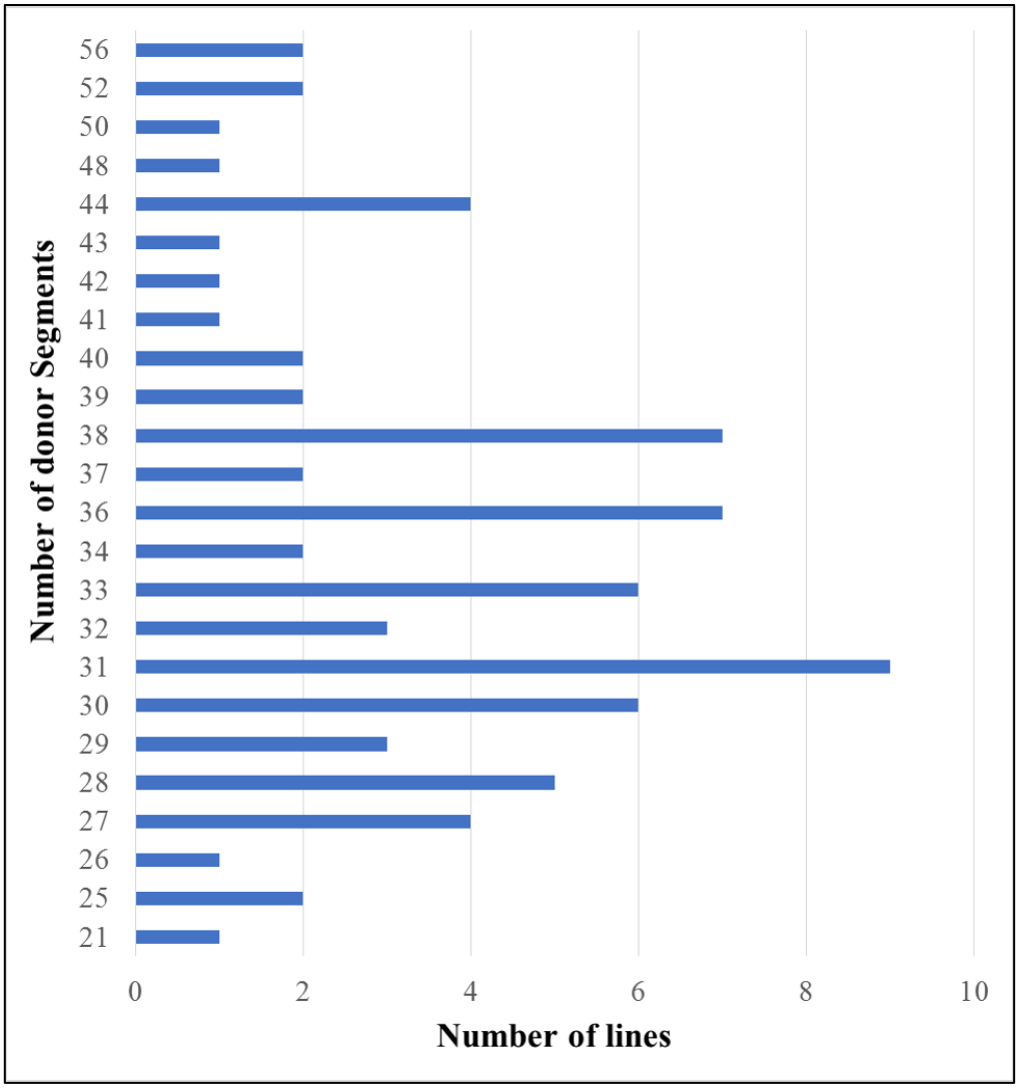

Figure 10: Distribution of the number of donor segments across the lines included in the introgression analysis.

The CSSL finder suggested the most suitable lines that have the introgressed segment sizes closest to the size of the desired segment of a particular chromosome (**Figure 11**). These lines are usually selected and subjected to subsequent foreground screening of the presence of the donor alleles for the chromosome under selection and

background screening for the presence of recurrent parent allele for the remaining chromosomes in the advanced backcrosses.

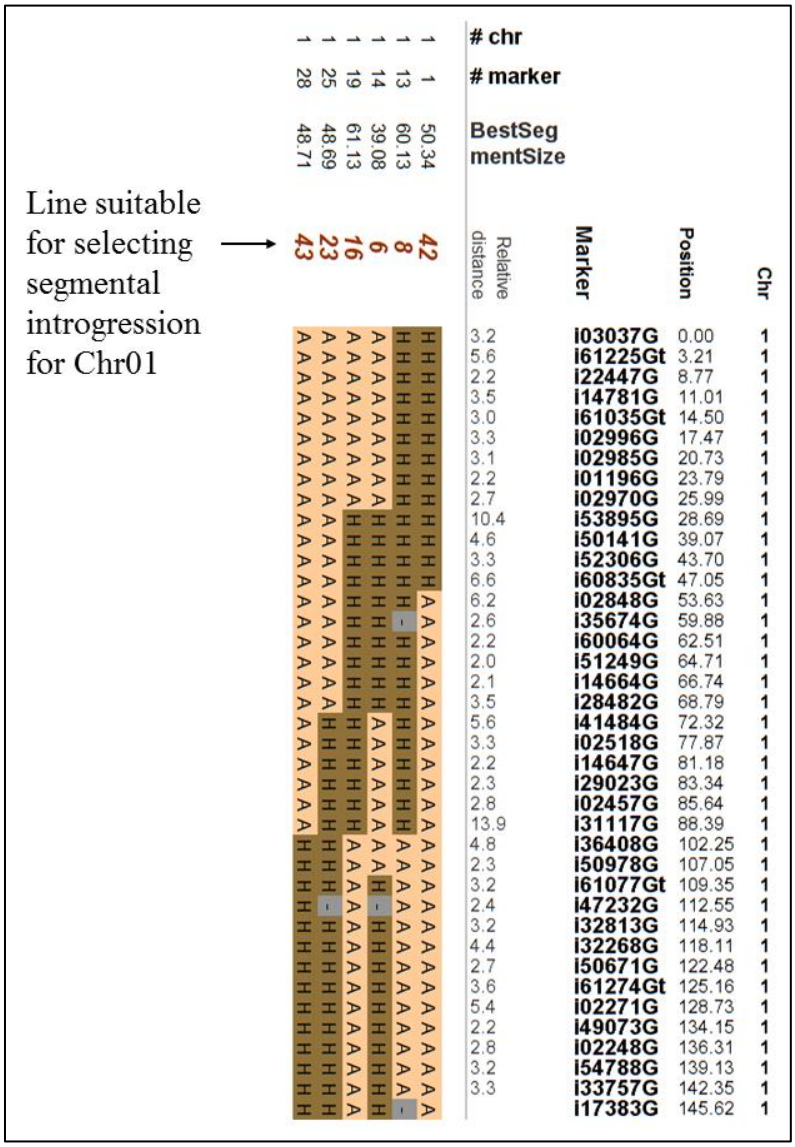

Figure 11: Graphical representation of lines suitable for obtaining desired segment sizes in the advanced backcrosses lines for Chr01. The lines 6, 8, 16, 23, 42 and 43 have segment introgression sizes ranging from 39 – 61 cM for Chr01 and are most suitable to select for overlapping segments of the desired size for Chr01 in the advanced backcrosses.

Simplex SNP assay validation panel

A random set of 52 SNP markers that were mapped to the linkage groups were selected at the rate of 2 markers per LG and were tested on the KASP assays to estimate their applicability through to the KASP platform for marker assisted selection. The results of the SNP validation panel indicated that the 83.6% (43/52) of the selected SNP resulted in scorable PCR assays. However, only 53.5% (23/43) of these markers produced "co-dominant" clusters where the heterozygous cluster was distinct from the homozygous recurrent parent cluster. The remaining 46.7% (20/43) of the markers produced "dominant" clusters meaning that the heterozygous clusters were indistinguishable from the recurrent parent clusters. The "failed" SNP markers produced random amplification patterns that did not result in any scorable patterns.

Defining half of the 52 SNPs as "A-genome SNPs" to denote that they mapped to AD-linkage groups 1 through 13, exactly 50% (13/26) exhibited "co-dominant" KASP assays, 38.5% (10/23) "dominant" KASP assays and 11.5% "failed" KASP assays (**Figure 12**). When these SNP sequences were subjected to BLASTn, 21 aligned to the BGI *G. arboreum* genome sequence available at https://www.cottongen.org/, and 7 aligned to the JGI *G. raimondii* (D_5) reference sequence. Two of the "co-dominant" SNP markers could not be aligned to the reference sequence. Similar statistics were also calculated for the analogous set of 26 "D-genome SNPs", i.e., SNPs that mapped to D-subgenome linkage groups Chr14 to Chr26. Overall, 38.5% (10/26) yielded "co-dominant" KASP assays, 38.5 (10/26) produced "dominant" KASP assays and 23%

(6/26) produced “failed” KASP assays. BLASTn performed for these sequences showed that 23 of these SNP sequences could be aligned to the JGI *G. raimondii* (D_5) reference sequence and 12 of these 23 could also be aligned to BGI *G. arboreum* genome sequence.

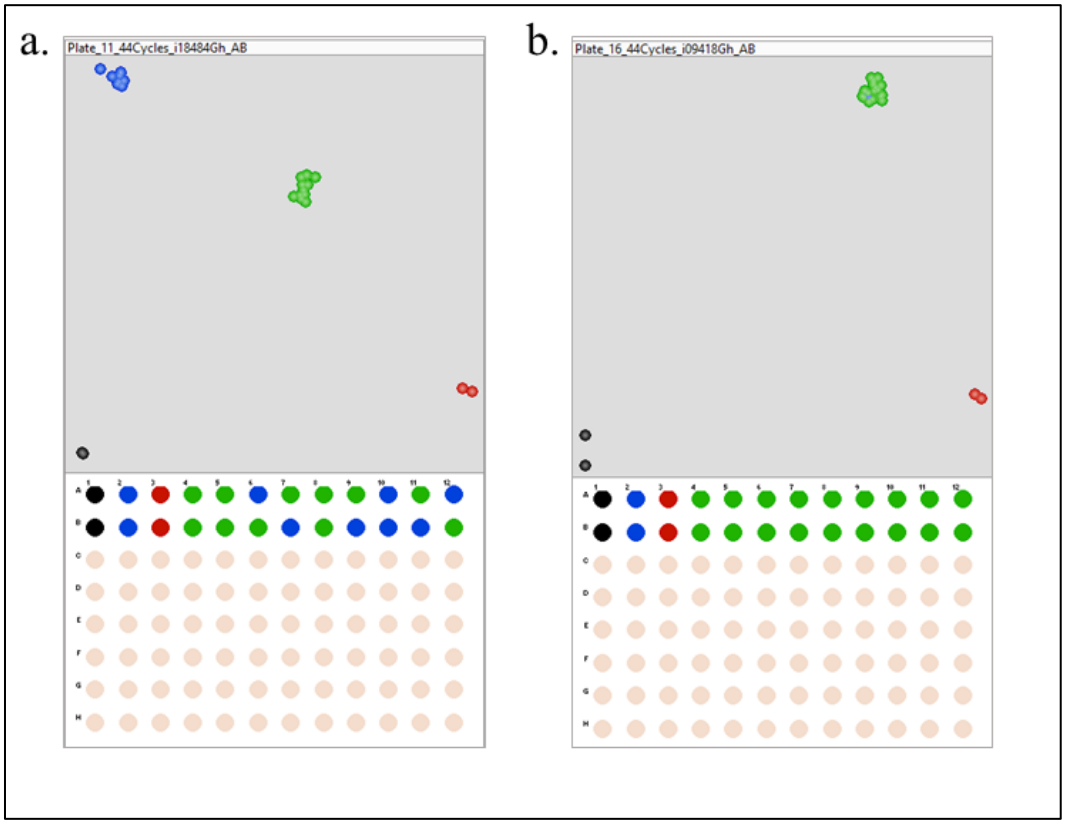

Figure 12: Examples of the results of "co-dominant" and "dominant" KASP assays for the SNPs tested in the validation panel. (**a.**) An example of a “co-dominant” SNP assay with clearly distinguishable clusters for the homozygous and heterozygous genotypes. 44.2 % of the tested SNP showed this pattern of clustering (**b.**) An example of “dominant” SNP assay. 38.4% of the tested SNP show this type of clustering pattern. Though this is an assayable SNP, it is not a useful SNP for marker-assisted breeding when the *G. hirsutum* SNP is dominant (as observed in this scenario), because the heterozygous cluster will be indistinguishable from one of the parental cluster.

The results of the SNP validation panel are summarized in the **Table 10**. The overall conversion rate can be summarized to 44% "co-dominant" SNP assays, 35.5% "dominant" SNP assays and 17.5% "failed" SNP assays.

Table 10: Results of SNP validation panel.

SNP ID	Chr	Remarks	Gr_Chr	Gr_pos	Ga_Chr	Ga_Pos
i52306Gb	AD01	Dominant	Chr02	59369280	Ca5	130908372
i30614Gh	AD01	Co-dominant	*	0	scaffold1123	293911
i03066Gh	AD02	Dominant	Chr03	85126	Ca1	137444500
i29065Gh	AD02	Dominant	*	0	Ca5	28894006
i23426Gh	AD03	Co-dominant	*	0	Ca13	8883516
i54909Gb	AD03	Dominant	*	0	*	0
i43499Gh	AD04	Co-dominant	Chr12	29491283	Ca2	106863526
i24385Gh	AD04	Co-dominant	*	0	*	0
i09277Gh	AD05	Dominant	Chr09	8974637	Ca6	87323212
i09151Gh	AD05	Failed	Chr09	6470665	Ca6	70332613
i32075Gh	AD06	Co-dominant	*	0	Ca4	114817439
i61821Gt	AD06	Co-dominant	*	0	Ca11	31348924
i18226Gh	AD07	Co-dominant	Chr01	14294029	*	0
i41382Gh	AD07	Co-dominant	*	0	Ca1	60150643
i49570Gh	AD08	Dominant	*	0	*	0
i30796Gh	AD08	Failed	*	0	Ca7	89120405
i49356Gh	AD09	Dominant	*	0	Ca10	94090367
i59188Gb	AD09	Dominant	*	0	Ca10	60942594
i29528Gh	AD10	Failed	*	0	Ca3	5157768
i00854Gh	AD10	Co-dominant	*	0	*	0
i07000Gh	AD11	Co-dominant	Chr07	7519312	Ca4	99827257
i65080Gm	AD11	Co-dominant	*	0	Ca3	80112376
i48341Gh	AD12	Co-dominant	*	0	Ca9	60296834
i52587Gb	AD12	Dominant	Chr08	35546676	Ca9	19771774
i61586Gt	AD13	Co-dominant	*	0	Ca5	143507574
i38406Gh	AD13	Dominant	*	0	Ca5	101971132
i56884Gb	AD14	Dominant	Chr05	38447335	*	0
i15353Gh	AD14	Dominant	Chr05	2714264	*	0
i02792Gh	AD15	Co-dominant	Chr02	55826232	Ca12	60934118
i18484Gh	AD15	Co-dominant	Chr02	55794466	Ca12	60985256

Table 10: Continued.

SNP ID	Chr	Remarks	Gr_Chr	Gr_pos	Ga_Chr	Ga_Pos
i01980Gh	AD16	Failed	Chr01	40478313	Ca1	94996193
i00934Gh	AD16	Failed	*	0	*	0
i37731Gh	AD17	Co-dominant	Chr03	19923340	*	0
i39720Gh	AD17	Dominant	*	0	*	0
i32363Gh	AD18	Co-dominant	Chr02	28870040	*	0
i63450Gm	AD18	Failed	Chr13	47382132	Ca5	28672710
i09418Gh	AD19	Dominant	Chr09	11941394	Ca6	52009896
i23637Gh	AD19	Co-dominant	Chr09	4464272	*	0
i56485Gb	AD20	Dominant	*	0	*	0
i12092Gh	AD20	Co-dominant	Chr11	50731042	Ca8	83571406
i51853Gb	AD21	Dominant	Chr07	25626972	Ca5	105880519
i07420Gh	AD21	Failed	Chr07	47416677	Ca3	127233934
i51144Gb	AD22	Co-dominant	Chr12	4176656	Ca2	49839013
i12867Gh	AD22	Dominant	Chr12	32873433	Ca2	90257931
i05827Gh	AD23	Dominant	Chr06	2788605	*	0
i34362Gh	AD23	Co-dominant	Chr06	47434015	*	0
i04583Gh	AD24	Co-dominant	Chr04	56356483	Ca7	89196097
i38741Gh	AD24	Failed	Chr04	16785724	*	0
i10927Gh	AD25	Failed	Chr10	12366120	Ca1	118231882
i17070Gh	AD25	Co-dominant	Chr10	155184	*	0
i56086Gb	AD26	Dominant	Chr08	16079117	*	0
i47939Gh	AD26	Dominant	Chr08	56585236	*	0

Discussion

This interspecific map is the first high-density linkage map of the interspecific hybrid of *G. hirsutum* and the A_2D_1 synthetic tetraploid with 26 linkage groups corresponding to the 26 allotetraploid chromosomes. In fact, it is the first high-density interspecific map between the cultivated *G. hirsutum* with any diploid species genome (https://www.cottongen.org/tools/cmap/viewer, accessed 2016 Oct. 09).

Implications of chromosomal rearrangements on linkage mapping

During the preliminary stages of the linkage map analysis, it was observed that SNP marker loci normally found in linkage groups corresponding to allotetraploid Chr01, Chr02 and Chr03, were grouped together with the default parameter, even at a LOD score of 19. A similar pattern was observed with Chr04 and Chr05. A previous study based on simulations found that the presence of reciprocal translocations tends to confound mapping software that detect "pseudo-linkage" between markers that are on different linkage groups, and leads to the formation of one large linkage group for the segments involved in the translocation (Livingstone, Churchill, & Jahn, 2000). Since it has been previously reported that the tetraploid A-genome of *G. hirsutum* and the two A-genome species differ by the presence of reciprocal translocations (Menzel & Brown, 1954), the observed patterns of grouping by JoinMap® were not surprising.

Cytological studies involving interspecific hybrids developed with Asiatic and American wild diploids with *G. hirsutum* led to the conclusion that the *G. hirsutum* - *G. herbaceum* differed by two reciprocal translocations, one involving tetraploid Chr02 and Chr03 and that presence of this translocation led to the formation of a tetravalent (dubbed "IV_1") in meiosis I of *G. hirsutum* - *G. herbaceum* interspecific hybrids (Menzel & Brown, 1954). A second translocation existed and dubbed IV_2. Menzel et al. (1985) mapped IV_1 breakpoints to near the centromeres of chromosomes 2 and 3, and those of IV_2 to near the centromeres of chromosomes 4 and 5. In addition, an additional

reciprocal translocation was found to distinguish the genomes of *G. herbaceum* and *G. arboreum*. Menzel et al. (1985) mapped IV_a breakpoints to near the centromeres of chromosomes 1 and 2. Moreover, it became clear that this additional translocation arose in *G. arboreum* lineage, because it further separated *G. hirsutum* and *G. arboreum*, too, which were found to differ by three reciprocal translocations involving the tetraploid A-subgenome Chr01 through Chr05. As a result, the metaphase chromosomes for interspecific hybrids from *G. hirsutum-G. arboreum* species have been shown to form a hexavalent (VI_1) involving Chr01, Chr02 and Chr03 and a tetravalent (form IV_2) involving Chr04 and Chr05 in addition to the 21 bivalents (II) (Menzel & Brown, 1954) (**Figure 13**).

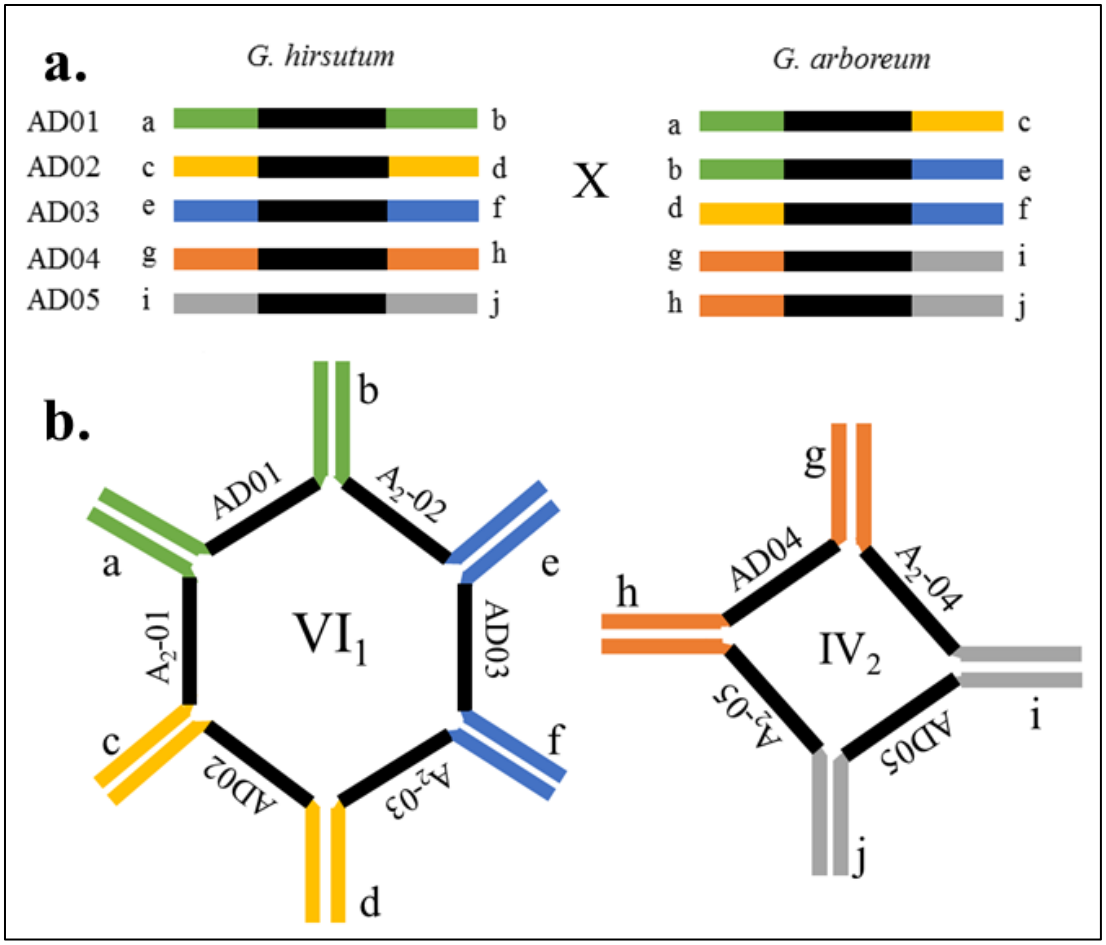

Figure 13: Stick diagrams of three chromosome translocations that distinguish the A-subgenome of *G. hirsutum* from the A_2 genome of *G. arboreum*. Redrawn from (Menzel & Brown, 1954). The interstitial segments (black) are each of unknown physical length but would include the low-recombination peri-centromeric regions of each chromosome. (a.) Relationships among distal segments (color-coded and labeled a-j) of chromosomes AD01-AD05 of *G. hirsutum* versus chromosomes of *G. arboreum*. (b.) Homologous pairing relationships of chromosome segments in meiotic multivalents in *G. hirsutum* and *G. arboreum* interspecific hybrids. These involve five chromosome pairs and three reciprocal translocations. Heterozygotes with a crossover in each distal segment form [LEFT] a closed hexavalent (VI_1) involving *G. hirsutum* chromosomes 1-2-3, due to a complex (double) translocation, and [RIGHT] a closed quadrivalent (IV_2) involving *G. hirsutum* chromosomes 4-5, due to single translocation.

Chromosome translocations complicate linkage map construction by causing co-recovery of markers from more than one chromosome, whereas recovery would normally be independent for loci of two non-homologous chromosomes. The availability of previously published high-density map from the interspecific F2 population from *G. hirsutum* x *G. barbadense* enabled alignment of SNP sequences of chromosomes AD01-AD05 to the *G. raimondii* D_5 sequence at high stringency (LOD 20.0). Given that the homology relationships between D_5 and both AD subgenomes, these alignments provided a basis for associating the SNP loci to the correct A-subgenome chromosomes of *G. hirsutum*. Re-estimation of the map distances with the regression mapping function and comparison of the map order to the *G. hirsutum – G. barbadense* interspecific map facilitated corrections of marker order and retention that upon subsequent linkage analysis yielded five individual linkage groups, one for each of these five chromosomes, and acceptable CheckMatrix plots of recombination intensities, as previously shown in Figure 8.

In a study of a barley population with a heterozygous reciprocal translocation, principal coordinate analysis (PCoA) was used to disentangle "pseudo-linkages" (Farré et al., 2011). In order to further investigate the applicability of this method in cotton, JoinMap® and CheckMatrix® were used to estimate the recombination frequencies of the combined group containing markers that were known to belong to three different chromosomes (Chr01, Chr02 and Chr03). A matrix was constructed for these recombination frequencies of the markers using R. Initially, a three-dimensional PcoA analysis was attempted for markers selected at every 10 cM along the linkage groups.

However, due to the presence of two translocations in the combined linkage group, the graphical results defied visual detection of a pattern that reflects the underlying translocation pattern. Subsequently, a modified approach was attempted with a total of 36 markers selected at the ends and the hypothetical peri-centromeric regions (around the mid-point) of the independent linkage groups. The length of the regions selected varied from 6 – 18 cM (which corresponded to ~30 cM in the preliminary map). Height plots were drawn from the matrix for the selected markers, and it was observed that at a height of 0.15, the data separated into distinct clusters. A hierarchical clustering plot then drawn for the selected data, when cut at 0.15, seven distinct branches were observed. When these branches were labelled with the chromosome and position of markers in individual linkage groups, a pattern was observed wherein each branch corresponded to different regions of the hexavalent (**Figure 14**).

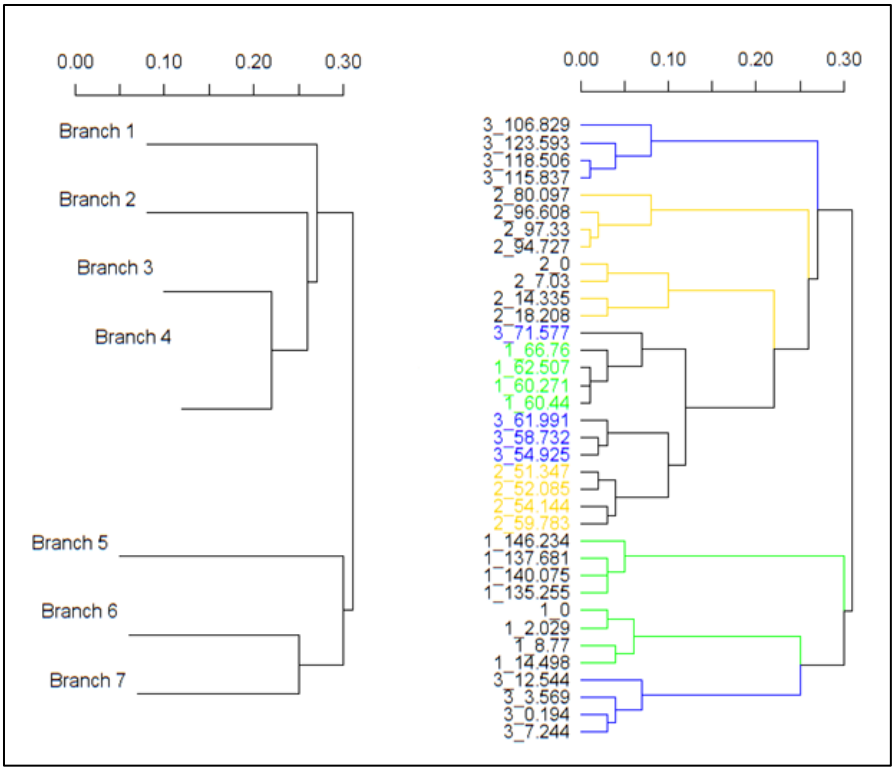

Figure 14: Hierarchical clustering patterns observed in the selected markers from the chromosomes involved in the hexavalent formation. (Left) Hierarchical tree of reduced complexity showing just the seven distinct branches observed when the tree is separated at h=0.15 (arbitrary value of similarity between the selected markers) (Right) Complete hierarchical clustering (no restriction on h-value) showing the corresponding chromosome number and positions of the selected markers in the linkage groups known to be involved in a formation of the hexavalent due to the presence of two reciprocal translocations.

These clusters were compared to the diagrammatic representation of the hexavalent, and it was observed that 6 of the 7 branches corresponded to ends of the 3 chromosomes. The remaining branch corresponded to markers selected from the possibly peri-centromeric regions of the three chromosomes (**Figure 15**). This pattern suggests that there is little or no recombination represented among BC1F1 progeny among parental sets of these loci and that they are co-segregating haplotypically due to

‘pseudo-linkage’. Contributing factors could be patterns of recombination, patterns of meiotic disjunction, sexual transmission and/or viability. We hypothesize that the central regions of these chromosomes are partially to largely pericentromeric and thus have low rates of recombination; if they contain critical genes, then the viability or relative health of gametophytes, especially pollen, and zygotes may hinge on the presence of balanced (parental) sets of these regions may be crucial. Additional assessments of the analytical methods and the explanatory hypothesis seem warranted.

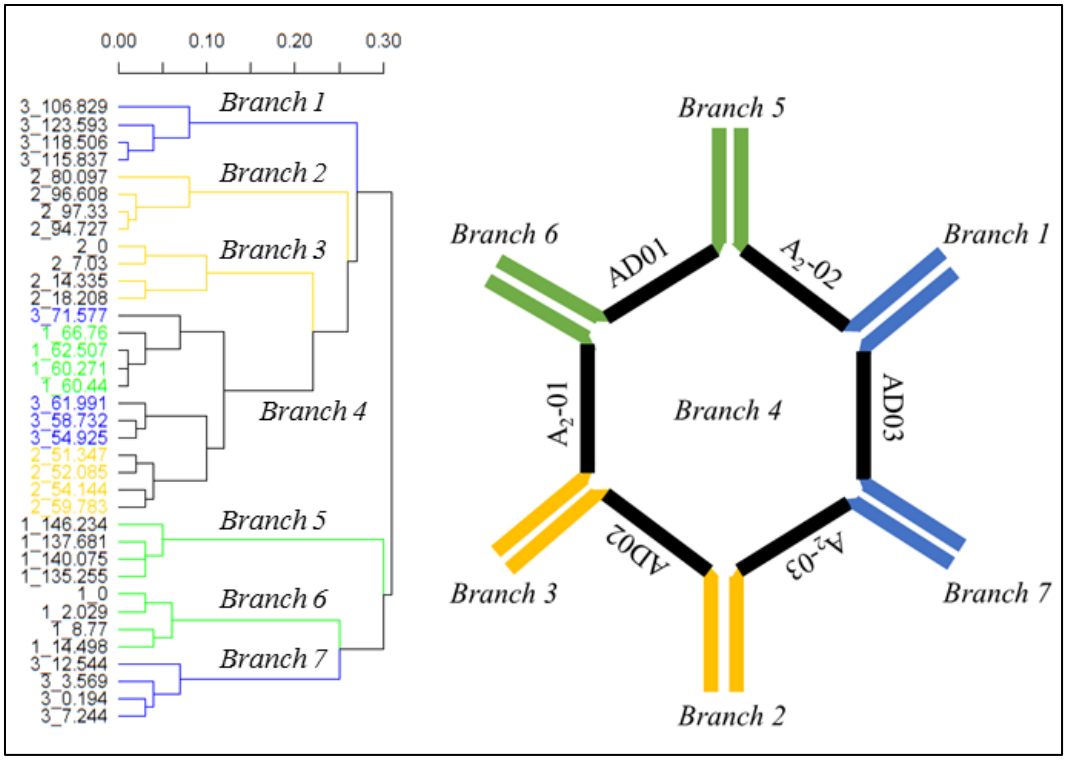

Figure 15: Relationship between multivariate analysis of SNP locus inheritance among the A_2D_1 / *G. hirsutum* BC1F1 hybrids and a stick model of key segments in complex reciprocal translocation hexavalents (VI) of the A_2D_1 / *G. hirsutum* F1 hybrid parent. *Right:* The hexavalent stick diagram depicts seven modeled segments, where the black central segment of each chromosome represents the pericentromeric region, flanked in each case by two colored segments with relatively higher rates of recombination, one segment distal the respective breakpoint, and one colored segment opposite the respective breakpoint. *Left:* The dendogram depicts results from hierachical cluster analysis of four SNP markers per segment of the VI. As can be seed from the identify of each marker (chromosome_cM map-position), each cluster branch identifies a genomic region that corresponds to the relative genetic positions in the linkage groups involved in its formation. Co-inheritances among 3x4 markers of the three pairs of non-homologous central regions are more closely related than are those of 6x4 markers of distal regions at opposite sides of the same chromosomes.

Linkage map analysis

Linkage maps for the linkage groups involved in the reciprocal translocation and those that had discordant 2D plots drawn using CheckMatrix were constructed using the frame work of shared markers from the *G. hirsutum* - *G. barbadense* interspecific map (Hulse-Kemp et al., 2015).

For the linkage groups that corresponded to A-genome chromosomes that were involved in the reciprocal translocations, the groups were selected at a LOD of 20, such that the individual groups corresponded to contiguous segments of the interspecific map of *G. hirsutum* - *G. barbadense*. The genotype data for these selected groups were pooled and then uploaded as individual projects that corresponded to each allotetraploid chromosome in JoinMap®. Since the groups were selected at a LOD score higher than 15 in the initial project, the default parameters for the chromosome specific JoinMap projects were altered such that all loci included will be forced to group together. The map distances for the complete group were then estimated using the regression mapping function. The 2D plots were generated using CheckMatrix to verify the map order and the final map was selected.

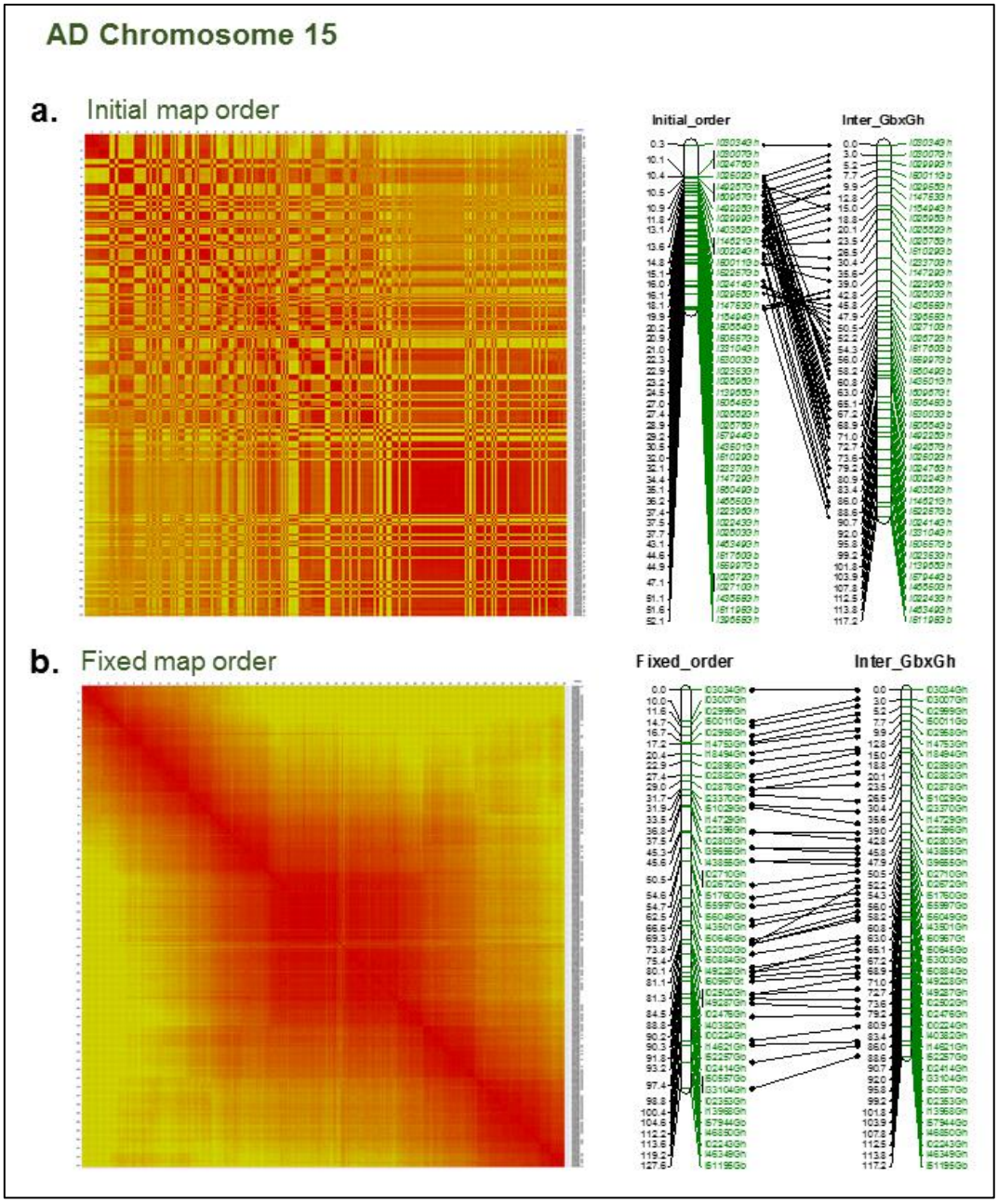

Figure 16: Analysis linkage maps of AD chromosome 15 before (a) and after (b) correction. (a) CheckMatrix 2D plot of linkage relationships for the *de novo* assembly and the linkage map connecting homologous markers for linkage group corresponding to AD chromosome 15 for the initial *de novo* map order to reference high-density linkage map of Hulse-Kemp (2015). (b) Re-calculated 2D plot and linkage map showing homologous markers between two maps for the corrected map order.

For those linkage groups that had discordant 2D plots, Chr05, Ch12, Chr15, Chr16, Chr20, and Chr24, the preliminary maps were compared to the interspecific map of *G. hirsutum* - *G. barbadense* (Hulse-Kemp et al., 2015). It was observed that the discordance in the 2D plots corresponded to the regions with significant deviations in the map order between the two maps (**Figure 16**). Therefore the map order of the *G. hirsutum* - *G. barbadense* interspecific map were used as the suggested start order for these discordant linkage groups and the map distances were re-estimated. In addition to the improvement in the 2D plots, the calculated lengths of the linkage groups increased and was closer to the estimates calculated from the *G. hirsutum* - *G. barbadense* interspecific map. The perceived amount of discordance seemed proportional to the increase in the map length of the corrected order of the linkage groups.

Table 11: Summary of the map order correction for the discordant linkage groups.

Chromosome	Total loci	Map length (cM)			
		Before correction	After correction	Interspecific GhxGb	% Change
AD05	723	94.4	177.41	210.4	46.8
AD12	586	122.5	150.151	151.5	18.4
AD15	608	56.6	127.578	117.2	55.6
AD16	620	89.3	125.532	133.5	28.9
AD20	605	80.8	100.47	151.7	19.6
AD24	778	84.6	109.097	129.5	22.5

Comparisons with the interspecific linkage group maps from *G. hirsutum* and *G. barbadense* F2 analysis revealed that the new A_2D_1-BC1F1 maps align well and shared a total of 10,902 mapped loci. Comparison of individual chromosomes indicated that an

average of 57.1% of the mapped A-subgenome markers were shared, while 55.6% of D-subgenome markers were shared. There were observed regions with differences in map order (**Figure 17**). A chromosome-by-chromosome comparison between the two interspecific maps was made for the numbers of shared loci and map lengths (**Table 12**).

Table 12: Chromosome-by-chromosome comparison of the interspecific A_2D_1-BC1F1 map with the interspecific GhxGb F2 map.

Chromosome	A_2D_1 map		Interspecific GhxGb		Shared loci
	Total loci	Map length	Total loci	Map length	
AD01	530	147.2	668	144.3	419
AD02	225	100.7	499	132.7	173
AD03	381	124.0	618	138.7	314
AD04	159	85.7	362	109.2	122
AD05	723	177.4	1038	210.4	616
AD06	449	128.2	590	136.6	360
AD07	511	118.9	648	150.2	402
AD08	957	122.9	1119	171	757
AD09	492	112.4	595	145.6	398
AD10	521	139.1	736	155.9	398
AD11	650	162.1	840	191.7	527
AD12	586	150.1	745	151.5	476
AD13	673	133.9	872	148.2	547
AD14	496	128.9	814	160.4	309
AD15	608	127.6	748	117.2	405
AD16	620	125.5	756	133.5	467
AD17	421	83.8	504	107.3	293
AD18	553	119.5	647	125.6	383
AD19	975	124.0	1680	225.4	776
AD20	605	100.5	647	151.7	433
AD21	422	86.6	559	184.4	258
AD22	311	93.5	447	128.1	220
AD23	483	106.5	594	127.1	372
AD24	778	109.1	941	129.5	541
AD25	661	123.5	839	130.9	507
AD26	621	139.1	685	147.4	429
Total:	**14411**	**3170.3**	**19191**	**3854.5**	**10902**

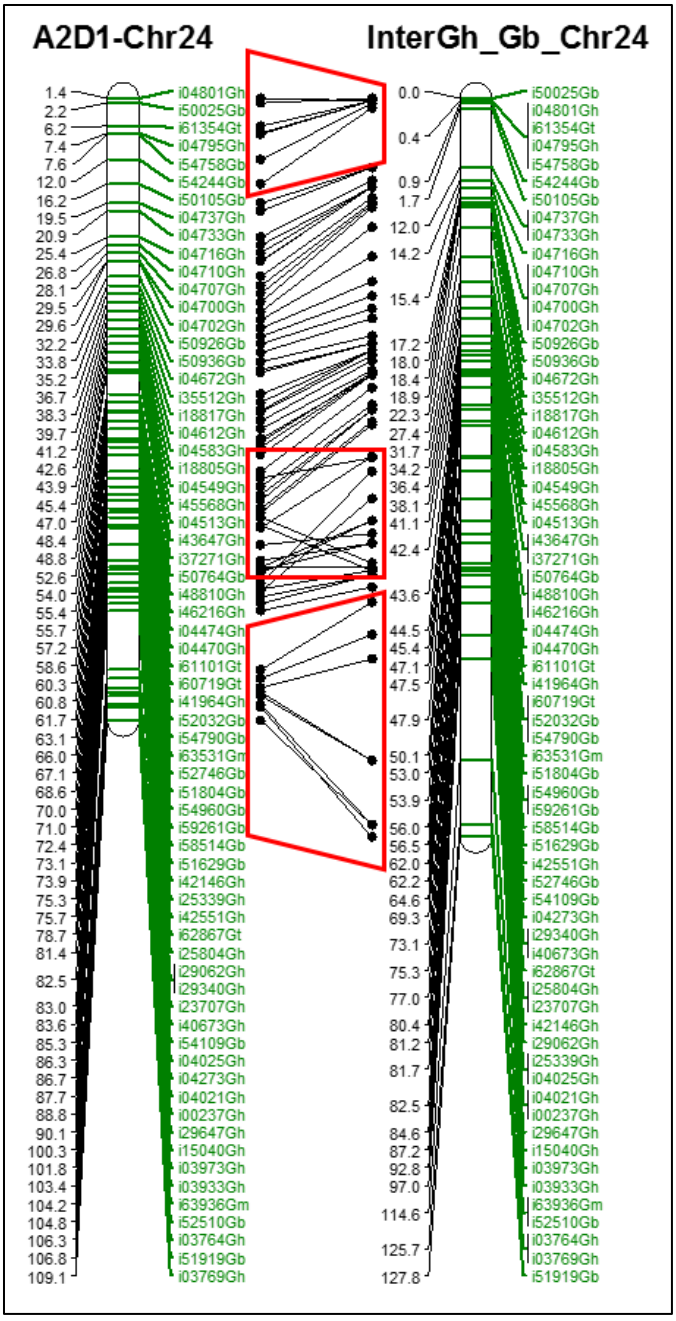

Figure 17: The *de novo* map order for Chr24 was improved with the suggested start order from the interspecific map of *G. hirsutum* - *G. barbadense* (Hulse-Kemp et al., 2015). The final map order was compared back to the interspecific map. The red trapezoidal regions highlight the observed changes in the relative distance or the order of markers in the corresponding linkage groups. However, since the 2D plots were acceptable, there were no indications that the A_2D_1 map order was spurious.

The difference in overall map length between the maps was 684.2 cM, which corresponds to a 17.8% decrease. With an exception of A-subgenome Chr01 and the corresponding D-subgenome homeolog Chr15, there was an average 14.3% decrease in the A-subgenome chromosomes (excluding the chromosomes involved in the translocations) and an average 18.9% decrease in the D-subgenome chromosome (**Figure 18**). The observed higher percentage of decrease in the D-subgenome chromosomes can possibly be due to the greater divergence between the *G. thurberi* (D_5) genome and the D-subgenome contributor of the allotetraploid species (Paterson et al., 2012; Jonathan F Wendel & Cronn, 2003).

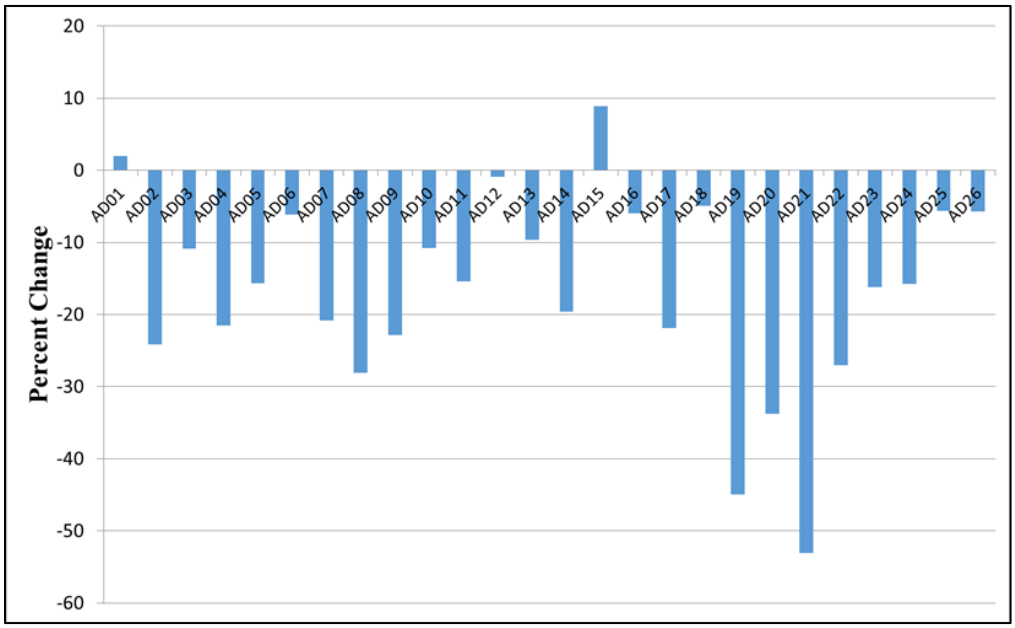

Figure 18: Graph showing the percent change in the linkage group length between the A_2D_1-BC1F1 map with the interspecific *G. hirsutum* - *G. barbadense* F2 map (Hulse-Kemp, 2015). It can be observed that, with an exception for Chr01 and its homeolog Chr15, there is trend of decrease in map length. It was an average 14.3% decrease for the A-subgenome and 18.9% decrease for the D-subgenome.

The final linkage groups were further classified into 2,692 recombination bins. Each recombination bin includes all markers within regions where in the individuals included in the population did not show any changes in genotype due to recombination. The advantage of binning markers is that it allows eliminating the errors in relative genetic distance calculated due to missing data. Expectedly, there were relatively higher numbers of markers mapped to bins that possibly correspond to the hypothesized peri-centromeric regions of the chromosomes (**Figure 19**).

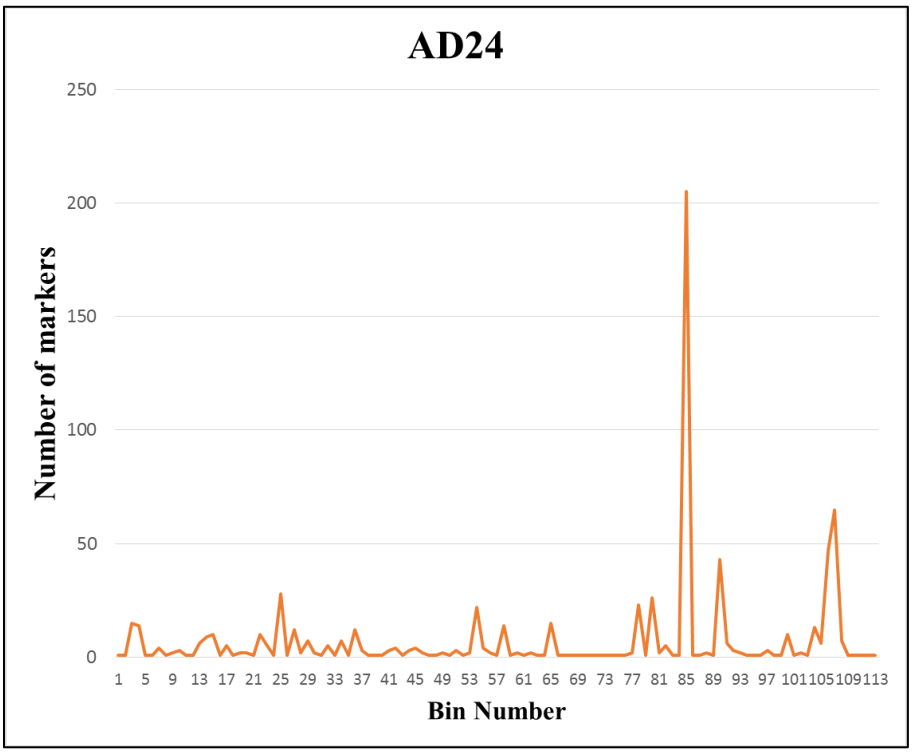

Figure 19: Graph showing the relative number of marker mapped to each bin for Chr24. The peak observed at bin 85 corresponds to the 205 markers mapped to the region of 83 – 84 cM for that linkage group. The markers mapped to bin 85 accounts for 26% of the markers mapped to the linkage group. It is reasonable to infer that this region corresponds to or is located within the pericentromeric heterochromatin of chromosome-24.

The information available from the linkage map, when used in conjunction with a simplex genotyping pipeline, e.g., high-throughput DNA extraction and KASP assays, could significantly reduce the cost and time required for marker-assisted selection of large backcross introgression population and the subsequent construction of Introgression Lines (ILs) for the involved A_2 and D_1 genomes. The map will also enable to improve the genome assembly efforts of the A_2 and D_1 genomes. In addition to aiding in the construction of the CSSLs, the high-density linkage map and validated KASP assays will be important for their utilization in breeding programs. Previous work done in the Stelly lab involving wide-cross introgression breeding has identified that intercrosses when used in combination with marker-assisted selection are better suited for recovering rare recombination events within the region of interest (Zheng et al., 2016). Intercrosses are crosses between introgression lines that have overlapping segments in the region of interest. It can be inferred that marker-assisted selection will not just aid introgression, but will be crucial to for analytical purposes as well as for selection and recovery of targeted recombination products in regions of interest and to select multilocus genetic combinations for pyramiding desirable genes and eliminating deleterious one (**Figure 20**).

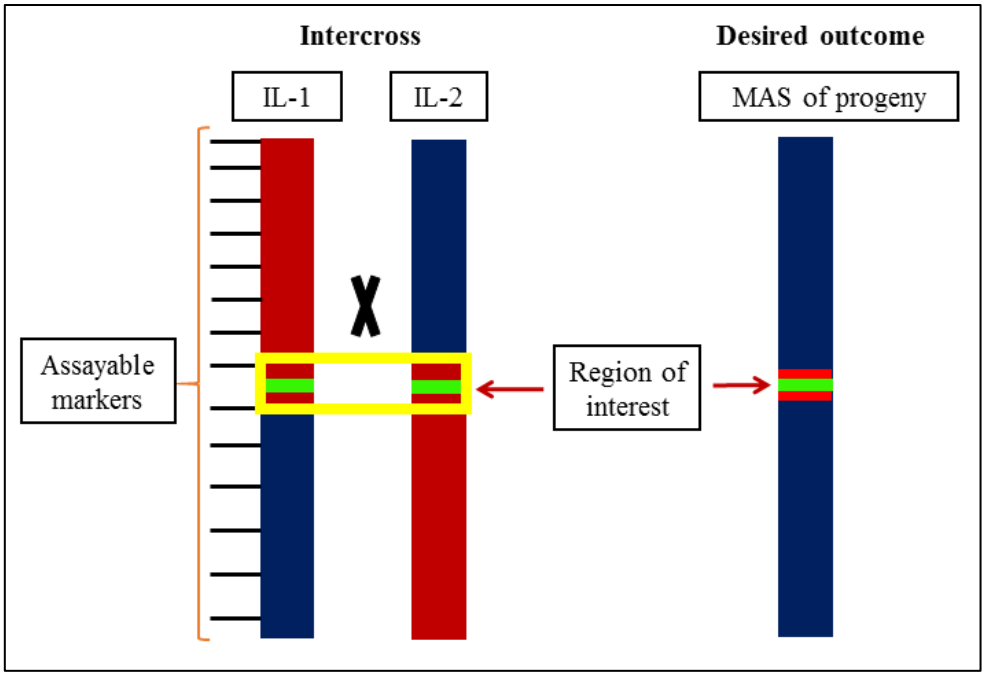

Figure 20: Potential use of simplex genotyping in combination with the Chromosome Segment Substitution Lines in downstream applied breeding. Availability of assayable markers along the chromosomes will allow for using intercrosses between introgression lines with the region of interest located in the overlapping segments and increasing the ability to recover a rare recombination event within the region of interest.

CHAPTER III

CONCLUSIONS

A customized "Gh-GP2 cluster file" was developed by re-assessing SNP signal distributions across a new screening panel to improve the call frequency and accuracy of automated genotyping of introgression populations developed using germplasm from the secondary gene pool, including Old World species (F_1 and A_2 genomes) and New World species (D_1 and $D_{2\text{-}1}$ genomes). The effort was successful in that the Gh-GP2 cluster file reduced the number of "zeroed" SNPs (<1% call frequency) by 2,536 and improved the call rate - increasing the number of SNPs having a call frequency >0.999 by 3,262. Additionally it was observed that the accuracy of the genotype call for the population was also improved and increased the total number of polymorphic SNPs for the A_2D_1-BC1F1 mapping population by 2,780 (20%). Thus, the customized cluster file provides a superior resource for automated genotype calling for populations containing introgressions from A_2 and D_1 germplasm. Some of these advantages will expectedly extend, by varying extents, to Upland cottons containing introgression of other diploid germplasm, especially from the secondary gene pool.

Using the CottonSNP6K array, 72 plants from the backcross population of A_2D_1 and Upland cotton *G. hirsutum* were automatically genotyped using the customized cluster file and a high-density linkage map of markers was constructed. This map comprises 14,411 SNP markers in 26 linkage groups that correspond to the 26 chromosomes. These linkage groups have a combined map length of 3,170 cM, and an

average density of ~4 markers per cM. The largest gap observed has a length of 14 cM. Linkage groups of A-subgenome chromosomes averaged 107 recombination bins per chromosome and an average density of 4.8 SNP markers per bin, whereas linkage groups from the D-subgenome averaged 100 recombination bins apiece and had an average density of 5.9 SNP markers per bin. The average length of A-subgenome linkage group maps (130.9 cM) was about 16% larger than the D-subgenome average (112.9 cM).

The rate at which mapped CottonSNP63K SNP markers could be converted to KASP assays indicated the rate at which good markers in the multiplex platform could be converted to simplex assays. Using a random set of 52 SNP mapped markers selected at the rate of 2 markers per LG and tested on the KASP assays, 44.2% (23/52) produced "co-dominant" assays, an optimal result -- where the heterozygous cluster of each SNP was distinct from the respective homozygous parental cluster. The remaining 55.7% (29/52) of the markers either produced "dominant" (20/52) or "failed" assays (9/52). The "dominant" assays are cluster patterns where the heterozygous clusters were indistinguishable from the recurrent parent clusters and the "failed" assays produced random amplification patterns that did not result in any scorable patterns. This conversion rate was found to be acceptable for genome-wide marker validation purposes. One or more sets of scorable markers spaced at approximately regular map intervals across each linkage group will be developed and used in downstream marker-assisted breeding, i.e., for the construction of Chromosome Segment Substitution Line construction (CSSLs/ILs) and/or intercrossing efforts for introgression of a small-sized segments associated with specific traits of interest.

Towards that end, concomitant progress was also made toward development of advanced generations of backcross populations, which are now at the BC_4F_1 seed generation. Individual seed or seedlings of these advanced lines can be screened with the genome-wide SNP genotyping systems, such as the CottonSNP63K and/or simplex or low-plex SNP genotyping systems, such as the KASP-enabled markers to select the plants and lines most suitable for constructing a set of Introgression Lines (ILs) that collectively provide comprehensive donor genome coverage, i.e., represented in an quasi-isogenic Upland cotton background.

The combination of the resources developed during the course of this project will play an important role in aiding the introgression breeding efforts that utilize the A_2 and D_1 genomes into Upland cotton (*G. hirsutum*).

REFERENCES

Beasley, J. (1942). Meiotic chromosome behavior in species, species hybrids, haploids, and induced polyploids of *Gossypium*. Genetics, 27(1), 25.

Bioversity. (2015). Financial Statements 2015 - for the year ended 31 December 2015 Auditor's report. http://www.bioversityinternational.org/fileadmin/user_upload/online_library/publications/pdfs/BioversityFinancialStatements2015_web_2076.pdf.

Brar, D., & Khush, G. (1997). Alien introgression in rice *Oryza:* From molecule to plant Springer (pp. 35-47).

Brubaker, C. L., & Wendel, J. F. (1994). Reevaluating the origin of domesticated cotton (*Gossypium hirsutum*; Malvaceae) using nuclear restriction fragment length polymorphisms (RFLPs). American Journal of Botany, 1309-1326.

Culp, T., & Harrell, D. (1973). Breeding methods for improving yield and fiber quality of upland cotton (*Gossypium hirsutum* L.). Crop Science, 13(6), 686-689.

Eathington, S. R., Crosbie, T. M., Edwards, M. D., Reiter, R. S., & Bull, J. K. (2007). Molecular markers in a commercial breeding program. Crop Science, 47(Supplement_3), S-154-S-163.

Ergon. (2008). Literature review and research evaluation relating to social impacts of global cotton production for ICAC expert panel on social, environmental and economic performance of cotton.

https://www.icac.org/seep/documents/reports/literature_review_july_2008.pdf, 80.

Eshed, Y., & Zamir, D. (1994). A genomic library of *Lycopersicon pennellii* in *L. esculentum*: A tool for fine mapping of genes. Euphytica, 79(3), 175-179. doi:10.1007/BF00022516

Eshed, Y., & Zamir, D. (1995). An introgression line population of *Lycopersicon pennellii* in the cultivated tomato enables the identification and fine mapping of yield-associated QTL. Genetics, 141(3), 1147-1162.

Fang, D. D., Hinze, L. L., Percy, R. G., Li, P., Deng, D., & Thyssen, G. (2013). A microsatellite-based genome-wide analysis of genetic diversity and linkage disequilibrium in Upland cotton (*Gossypium hirsutum* L.) cultivars from major cotton-growing countries. Euphytica, 191(3), 391-401.

FAO. (2010). The Second Report on the State of the World's Plant Genetic Resources for Food and Agriculture. http://www.fao.org/docrep/013/i1500e/i1500e.pdf, 399.

Farré, A., Benito, I. L., Cistué, L., De Jong, J., Romagosa, I., & Jansen, J. (2011). Linkage map construction involving a reciprocal translocation. Theoretical and Applied Genetics, 122(5), 1029-1037.

Fridman, E., Pleban, T., & Zamir, D. (2000). A recombination hotspot delimits a wild-species quantitative trait locus for tomato sugar content to 484 bp within an invertase gene. Proceedings of the National Academy of Sciences, 97(9), 4718-4723.

Fryxell, P. A. (1976). Germpool utilization: *Gossypium*, a case history: US Department of Agriculture, Agricultural Research Service, Southern Region.

Grover, C., Zhu, X., Grupp, K., Jareczek, J., Gallagher, J., Szadkowski, E., . . . Wendel, J. (2015). Molecular confirmation of species status for the allopolyploid cotton species, *Gossypium ekmanianum* Wittmack. Genetic Resources and Crop Evolution, 62(1), 103-114.

Grover, C. E., Gallagher, J. P., Jareczek, J. J., Page, J. T., Udall, J. A., Gore, M. A., & Wendel, J. F. (2015). Re-evaluating the phylogeny of allopolyploid *Gossypium* L. Molecular Phylogenetics and Evolution, 92, 45-52.

Gulati, A., & Turner, A. J. (1929). A note on the early history of cotton. Journal of the Textile Institute Transactions, 20(1), T1-T9.

Hadley, H., & Openshaw, S. (1980). Interspecific and intergeneric hybridization. Hybridization of Crop Plants, 133-159.

Harlan, J. R., & de Wet, J. M. (1971). Toward a rational classification of cultivated plants. Taxon, 509-517.

Hoisington, D., Khairallah, M., Reeves, T., Ribaut, J.-M., Skovmand, B., Taba, S., & Warburton, M. (1999). Plant genetic resources: What can they contribute toward increased crop productivity? Proceedings of the National Academy of Sciences, 96(11), 5937-5943.

Hulse-Kemp, A. M., Lemm, J., Plieske, J., Ashrafi, H., Buyyarapu, R., Fang, D. D., . . . Hinze, L. L. (2015). Development of a 63K SNP array for cotton and high-

density mapping of intra-and inter-specific populations of *Gossypium* spp. G3: Genes| Genomes| Genetics, 115.018416.

Hutchinson, J. (1962). The history and relationships of the world's cottons. Endeavour, 21, 5-15.

Hutchinson, J., & Dalziel, J. (1954). Flora of West Tropical Africa. Vol. 1, Pt. 1. Flora of West Tropical Africa. Vol. 1, Pt. 1.(Edn 2).

Hutchinson, J. B., Silow, R. A., Stephens, S. G.,. (1947). The evolution of *Gossypium* and the differentiation of the cultivated cottons. London: Oxford University Press.

Illumina, I. (2014). Infinium Genotyping Data Analysis http://www.illumina.com/Documents/products/technotes/technote_infinium_genotyping_data_analysis.pdf).

Innes, N. (1966). Inheritance of resistance to bacterial blight of cotton. III. *G. herbaceum* resistance transferred to tetraploid cotton. The Journal of Agricultural Science, 66(03), 433-439.

Iqbal, M., Reddy, O., El-Zik, K., & Pepper, A. (2001). A genetic bottleneck in the evolution under domestication of upland cotton *Gossypium hirsutum* L. examined using DNA fingerprinting. Theoretical and Applied Genetics, 103(4), 547-554.

Jeuken, M., & Lindhout, P. (2004). The development of lettuce backcross inbred lines (BILs) for exploitation of the *Lactuca saligna* (wild lettuce) germplasm. Theoretical and Applied Genetics, 109(2), 394-401.

Jia, Y., Jia, M. H., Wang, X., & Liu, G. (2012). Indica and japonica crosses resulting in linkage block and recombination suppression on rice chromosome 12. PLoS One, 7(8), e43066.

Kubo, T., Aida, Y., Nakamura, K., Tsunematsu, H., Doi, K., & Yoshimura, A. (2002). Reciprocal chromosome segment substitution series derived from japonica and indica cross of rice (*Oryza sativa* L.). Breeding Science, 52(4), 319-325. doi:10.1270/jsbbs.52.319

Kuspira, J., & Unrau, J. (1957). Genetic analyses of certain characters in common wheat using whole chromosome substitution lines. Canadian Journal of Plant Science, 37(3), 300-326.

Lacape, J.-M., Nguyen, T.-B., Thibivilliers, S., Bojinov, B., Courtois, B., Cantrell, R., . . . Hau, B. (2003). A combined RFLP SSR AFLP map of tetraploid cotton based on a *Gossypium hirsutum*× *Gossypium barbadense* backcross population. Genome, 46(4), 612-626.

Lander, E. S., & Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics, 121(1), 185-199.

Levi, A., Paterson, A. H., Barak, V., Yakir, D., Wang, B., Chee, P. W., & Saranga, Y. (2009). Field evaluation of cotton near-isogenic lines introgressed with QTLs for productivity and drought related traits. Molecular Breeding, 23(2), 179-195.

Lin, Z.-x., He, D., Zhang, X.-l., Nie, Y., Guo, X., Feng, C., & Stewart, J. M. (2005). Linkage map construction and mapping QTL for cotton fibre quality using SRAP, SSR and RAPD. Plant Breeding, 124(2), 180-187.

Livingstone, K., Churchill, G., & Jahn, M. (2000). Linkage mapping in populations with karyotypic rearrangements. Journal of Heredity, 91(6), 423-428.

Mammadov, J., Aggarwal, R., Buyyarapu, R., & Kumpatla, S. (2012). SNP markers and their impact on plant breeding. International Journal of Plant Genomics, 2012.

Mei, M., Syed, N., Gao, W., Thaxton, P., Smith, C., Stelly, D., & Chen, Z. (2004). Genetic mapping and QTL analysis of fiber-related traits in cotton (*Gossypium*). Theoretical and Applied Genetics, 108(2), 280-291.

Menzel, M. Y., & Brown, M. S. (1954). The significance of multivalent formation in three-species *Gossypium* hybrids. Genetics, 39(4), 546.

Mergeai, G. (2006). Introgressions interspécifiques chez le cotonnier. Cahiers Agricultures, 15(1), 135-143.

Meyer, J. R. (1957). Origin and inheritance of D2 smoothness in upland cotton. Journal of Heredity, 48(5), 249-250.

Moncada, P., Martinez, C., Borrero, J., Châtel, M., Gauch Jr, H., Guimaraes, E., . . . McCouch, S. R. (2001). Quantitative trait loci for yield and yield components in an *Oryza sativa*× *Oryza rufipogon* BC2F2 population evaluated in an upland environment. Theoretical and Applied Genetics, 102(1), 41-52.

NPGS. (1997). Cotton Germplasm Status Report. Retrieved from http://www.ars-grin.gov/npgs/cgc_reports/cotton97.html

Paterson, A. H., Wendel, J. F., Gundlach, H., Guo, H., Jenkins, J., Jin, D., . . . Udall, J. (2012). Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. Nature, 492(7429), 423-427.

Percival, A. E., & Kohel, R. J. (1990). Distribution, collection, and evaluation of *Gossypium*. Advances in Agronomy, 44, 225-256.

Rhyne, C. L., & Smith, F. H. (1965). Genetic aspects of gossypol content of leaves and flower buds of *Gossypium*. Crop Science, 5(5), 419-421.

Rick, C. M. (1974). High soluble-solids content in large-fruited tomato lines derived from a wild green-fruited species: University of California, Division of Agriculture and Natural Resources.

Ronen, G., Carmel-Goren, L., Zamir, D., & Hirschberg, J. (2000). An alternative pathway to β-carotene formation in plant chromoplasts discovered by map-based cloning of Beta and old-gold color mutations in tomato. Proceedings of the National Academy of Sciences, 97(20), 11102-11107.

Shen, X., Guo, W., Lu, Q., Zhu, X., Yuan, Y., & Zhang, T. (2007). Genetic mapping of quantitative trait loci for fiber quality and yield trait by RIL approach in Upland cotton. Euphytica, 155(3), 371-380.

Silow, R. (1944). The genetics of species development in the Old World cottons. Journal of Genetics, 46(1), 62-77.

Small, R. L., Ryburn, J. A., & Wendel, J. F. (1999). Low levels of nucleotide diversity at homoeologous *Adh* loci in allotetraploid cotton (*Gossypium* L.). Molecular Biology and Evolution, 16(4), 491-501.

Stephens, S. (1947). Cytogenetics of *Gossypium* and the problem of the origin of New World cottons. Advances in Genetics, 1, 431-442.

Stephens, S. (1958). Salt water tolerance of seeds of *Gossypium* species as a possible factor in seed dispersal. American Naturalist, 83-92.

Stewart, J. M., & Hsu, C. L. (1977). In-ovulo embryo culture and seedling development of cotton (*Gossypium hirsutum* L.). Planta, 137(2), 113-117.

Tanksley, S. D. (1988). Resolution of quantitative traits into mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. Nature, 335(721726), 6170.

Tanksley, S. D., & McCouch, S. R. (1997). Seed banks and molecular maps: unlocking genetic potential from the wild. Science, 277(5329), 1063-1066.

Ulloa, M. (2014). The diploid D genome cottons (*Gossypium* spp.) of the New World. World Cotton Germplasm Resources, 203.

Van Esbroeck, G., & Bowman, D. T. (1998). Cotton germplasm diversity and its importance to cultivar development. Journal of Cotton Science 2:121-129.

Van Ooijen, J. (2006). JoinMap® 4, Software for the calculation of genetic linkage maps in experimental populations. Kyazma BV, Wageningen, 33, 10.1371.

Vavilov, N. (1940). In The New Systematics, J. Huxley, Ed., Clarendon: Oxford.

Von Korff, M., Wang, H., Léon, J., & Pillen, K. (2004). Development of candidate introgression lines using an exotic barley accession (*Hordeum vulgare* ssp. *spontaneum*) as donor. Theoretical and Applied Genetics, 109(8), 1736-1745.

Voorrips, R. (2002). MapChart: software for the graphical presentation of linkage maps and QTLs. Journal of Heredity, 93(1), 77-78.

Wang, K., Song, X., Han, Z., Guo, W., John, Z. Y., Sun, J., . . . Zhang, T. (2006). Complete assignment of the chromosomes of *Gossypium hirsutum* L. by translocation and fluorescence *in situ* hybridization mapping. Theoretical and Applied Genetics, 113(1), 73-80.

Wang, P., Ding, Y., Lu, Q., Guo, W., & Zhang, T. (2008). Development of *Gossypium barbadense* chromosome segment substitution lines in the genetic standard line TM-1 of *Gossypium hirsutum*. Chinese Science Bulletin, 53(10), 1512-1517.

Weeden, N. F., & Wendel, J. F. (1989). Genetics of plant isozymes: Isozymes in plant biology. Springer (pp. 46-72).

Wendel, J. F. (1989). New World tetraploid cottons contain Old World cytoplasm. Proceedings of the National Academy of Sciences, 86(11), 4132-4136.

Wendel, J. F., Brubaker, C., Alvarez, I., Cronn, R., & Stewart, J. M. (2009). Evolution and natural history of the cotton genus. Genetics and Genomics of Cotton (pp. 3-22): Springer.

Wendel, J. F., Brubaker, C. L., & Percival, A. E. (1992). Genetic diversity in *Gossypium hirsutum* and the origin of upland cotton. American Journal of Botany, 1291-1310.

Wendel, J. F., & Cronn, R. C. (2003). Polyploidy and the evolutionary history of cotton. Advances in Agronomy, 78, 139-186.

Wendel, J. F., & Grover, C. E. (2015). Taxonomy and evolution of the cotton genus, *Gossypium*. Cotton (pp. 25-44). doi:10.2134/agronmonogr57.2013.0020

Young, N. D. (1999). A cautiously optimistic vision for marker-assisted breeding. Molecular Breeding, 5(6), 505-510.

Zamir, D. (2001). Improving plant breeding with exotic genetic libraries. Nature Reviews Genetics, 2(12), 983-989.

Zhang, Z., Rong, J., Waghmare, V. N., Chee, P. W., May, O. L., Wright, R. J., . . . Paterson, A. H. (2011). QTL alleles for improved fiber quality from a wild Hawaiian cotton, *Gossypium tomentosum*. Theoretical and Applied Genetics, 123(7), 1075-1088.

Zheng, X., Hoegenauer, K. A., Maeda, A. B., Wang, F., Stelly, D. M., Nichols, R. L., & Jones, D. C. (2015). Non-destructive high-throughput DNA extraction and genotyping methods for cotton seeds and seedlings. BioTechniques, 58(5), 234.

Zheng, X., Hoegenauer, K. A., Quintana, J., Bell, A. A., Hulse-Kemp, A. M., Nichols, R. L., & Stelly, D. M. (2016). SNP-Based MAS in Cotton under depressed-recombination for– flanking recombinants: Results and inferences on wide-cross breeding strategies. Crop Science. doi: 10.2135/cropsci2015.07.0436