DEVELOPMENT OF GENOMIC MARKERS AND MAPPING TOOLS FOR ASSEMBLING THE ALLOTETRAPLOID *GOSSYPIUM HIRSUTUM* L. DRAFT GENOME SEQUENCE

A Dissertation

by

AMANDA MARIE HULSE-KEMP

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

DOCTOR OF PHILOSOPHY

Chair of Committee,	David M. Stelly
Committee Members,	Alan E. Pepper
	Clare A. Gill
	James J. Cai
Intercollegiate Faculty	
Chair,	Craig Coates

May 2015

Major Subject: Genetics

ABSTRACT

Cotton (*Gossypium spp.*) is the largest producer of natural textile fibers. Most worldwide and domestic cotton fiber production is based on cultivars of *G. hirsutum* L., an allotetraploid. Genetic improvement of cotton remains constrained by alarmingly low levels of genetic diversity, inadequate genomic tools for genetic analysis and manipulation, and the difficulty of effectively harnessing the vastly greater genetic diversity harbored by other *Gossypium* species. Development of large numbers of single nucleotide polymorphisms (SNPs) for use in intraspecific and interspecific populations will allow for cotton germplasm diversity characterization, high-throughput genotyping, marker-assisted breeding, germplasm introgression of advantageous traits from wild species, and high-density genetic mapping. My research has been focused on utilizing next generation sequencing data for intraspecific and interspecific SNP marker development, validation, and creation of high-throughput genotyping methods to advance cotton research.

I used transcriptome sequencing to develop and map the first gene-associated SNPs for five species, *G. barbadense* (Pima cotton), *G. tomentosum*, *G. mustelinum*, *G. armourianum*, and *G. longicalyx*. A total of 62,832 non-redundant SNPs were developed. These can be utilized for interspecific germplasm introgression into cultivated *G. hirsutum*, as well as for subsequent genetic analysis and manipulation. To create SNP-based resources for integrated physical mapping, I used BAC-end sequences (BESs) and resequecing data for 12 *G. hirsutum* lines, a Pima line and *G. longicalyx* to derive 132,262 intraspecific and 693,769 interspecific SNPs located in BESs. These SNP data sets were used to help build the first high-throughput genotyping array for cotton, the CottonSNP63K, which now provides a standardized platform for global cotton research. I applied the array to two F_2 populations and produced the first two high-density SNP maps for cotton, one intraspecific and one interspecific. By resequencing two interspecific F_1 hypo-aneuploids, I also demonstrated that the chromosome-wide changes in SNP genotypes enable highly effective mass-localization of BACs to individual cotton chromosomes. These efforts provide additional validation and placement methods that can be directly integrated with the physical map being constructed for *G. hirsutum* and enable the production of a high-quality draft genome sequence for cultivated cotton.

DEDICATION

I dedicate my dissertation work to all those who have inspired, encouraged, and supported me throughout my scientific journey. Particularly to my loving parents Kathleen and Michael Hulse, who instilled in me a passion for science and encouraged me to follow my dreams; and to my unbelievably loving, patient, and supportive husband, Gabriel Kemp, without whom this work would not have been possible.

ACKNOWLEDGEMENTS

A special appreciation to my advisor, Dr. David Stelly, for providing guidance during my research and providing many opportunities to interact with the domestic and international research communities along the way. Thank you for teaching me to think critically throughout my project, allowing me to interact with numerous collaborators, and giving me the opportunity to present my research on the international level. Also to those distinguished faculty members who not only served on my committee but assisted me when needed on my projects: Dr. Alan Pepper, Dr. Clare Gill, and Dr. James Cai. A special thanks to Dr. Jeffrey Chen for being an external committee member. Thank you as well to my close collaborators, Dr. Allen Van Deynze and Dr. Hamid Ashrafi.

I am very thankful to those who have helped to take care of my plants in the greenhouse and the field, as well as planting and harvesting seed: Wayne Raska, Dr. Robert Vaughn, Dustin Kelley and numerous student workers. Thanks to all the members of the Stelly lab for encouragement and friendship along the way, particularly Dr. Xiuting Zheng, Andrea Maeda, Dr. Bo Liu, Kevin Hoegenauer, and Jiale Xu.

As a part of my work, I utilized many facilities on the Texas A&M University campus and am thankful to those departments and institutions for supporting my work. The Soil and Crop Science Department and Dr. Fei Wang for DNA extraction and genotyping work in the AgriGenomics Laboratory. The Texas A&M Institute for Genome Sciences and Society (formerly the Whole Systems Genomics Institute) and Dr. Penny Riggs, Kelli Kochan, Kranti Konganti and Dr. Andrew Hillhouse for their support

utilizing high-performance computing resources, running large numbers of cotton SNP chips, and high-throughput genotyping.

The development of the CottonSNP63K array was a collaborative effort of many groups and individuals; I would like to acknowledge the important contributions that led to the development of the cotton array: Dr. Iain Wilson and his CSIRO group for contributing large numbers of markers and samples, Dr. Martin Ganal, Jana Lemm and Joerg Plieske at TraitGenetics for assisting with cluster file development, Dr. Cindy Lawley at Illumina for assistance with development and technical assistance, Dr. David Stelly, Dr. Allen Van Deynze and myself for initiating the vision for developing the collaboration and seeing it through completion, and the other many contributors of germplasm and/or marker data sets.

I thank Dr. Jeffrey Chen for leading the National Science Foundation project which provided my funding as well as the other collaborators on the project, Dr. David Stelly, Dr. Brian Scheffler, Dr. Christopher Saski, Dr. Candace Haigler, Dr. Jeremy Schmutz, and Dr. Xueying Guan. Other funding was provided by Cotton Incorporated.

I would like to go on to thank a few fellow Genetics graduate students for their support from the very first day of our program, including Rachel Jordan, Stephen Caster, Jaime Brannan, Kylee Veazey, Roel Lerma, Lauren Dobson and Yvette Halley. Thank you for being a sounding board, providing constructive thoughts on projects and being stress relievers and making me laugh when needed.

Without my husband, Gabriel Kemp, I know I would not have been able to complete this work. Thank you for listening to my research problems even if you do not know the technicalities, the many times you've helped figure out how to make excel do what I wanted, for making sure that I ate dinner, and for taking softball breaks. Thank you for helping me to keep everything in perspective and to remember the important things in life. Also a special thanks to our three dogs, Rainy, Krieger, and Archer, who have diligently stayed at my side while I was writing or analyzing.

Through this journey I have met many talented and compassionate individuals who have unselfishly donated their time to help me along the way. I hope to not miss mentioning anyone's name, so I will simply say "Thank you ALL for being there for me!"

TABLE OF CONTENTS

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 that is produced in over 75 countries. While it is best known for the production of fiber for use in textiles, it is also a significant producer of cottonseed which can be used as feed for livestock or to produce cottonseed oil which is an important vegetable oil used in the human food industry. The world-wide cotton production generated over 118 million bales of fiber in 2013 (National Cotton Council, www.cotton.org). The United States contributed 12.9 million bales which was valued at $5.2 billion and had a direct economic impact on the US economy of $27.6 billion (US Department of Agriculture; National Agriculture Statistics Service). The yearly direct economic impact world-wide is estimated to be approximately $500 billion. Cotton production is especially important for many developing countries such as Uzbekistan and Pakistan, where cotton exports represent a significant portion of their exports.

World-wide cotton production stems from cultivation of four cotton species, two diploids (2*n*=2*x*=26) and two allotetraploids (2*n*=4*x*=52). Over 95% of the world crop is produced from cultivation of allotetraploid Upland cotton, or *Gossypium hirsutum* L. The remaining ~5% of the cotton produced is contributed mostly from Pima cotton, or *G. barbadense* L., and lesser contributions from A-genome diploids *G. arboreum* L. and *G. herbaceum* L. Upland cotton is the primary cultivated species as it generates a highyielding and thus high-profit product for farmers. Pima cotton is grown for its superior fiber qualities, however the increase in quality typically comes at the detriment to yield production. A-genome diploids are primarily only cultivated in small amounts in primarily Asia and Australia.

The *Gossypium* genus has been recently updated to contain 52 species, including 7 allotetraploid species and 45 diploid species (Wendel, Brubaker et al. 2009). Two allotetraploids that were previously associated as Upland types have recently been identified to represent distinct species, *G. ekmanianum* Wittmack $[AD]_6$ (Grover, Zhu et al. 2014) and $[AD]_7$ (Wendel - unpublished). The diploid species fall into eight genomic groups (A-G & K) that have adapted in different regions of the world. African and Asian cottons include A-genome diploids in which spinnable fiber initially appeared (Applequist, Cronn et al. 2001) as well as the B, E, and F genomes. The D-genome diploid clade, or new world diploids are found in Central and Northern America. The remaining diploid genomes (C, G, and K) are located in Australia.

Genomes of allotetraploids, designated as $[AD]_n$, arose evolutionarily from a single polyploidization event 1-2 million years ago (mya) near present day Mexico, between a new world D-genome diploid and an old world A-genome diploid (Wendel and Cronn 2003). The D-genome diploid ancestor has been determined to most closely related to the extant species *G. raimondii* Ulbrich, whereas the A-genome ancestor is closely related to the extant species *G. arboreum* L.. The phylogeny of the cotton lineage and divergence time between the A-genome and D-genome diploid lineages has been experimentally determined to be approximately 5-10 mya (**Figure 1.1**) (Wendel and Cronn 2003).

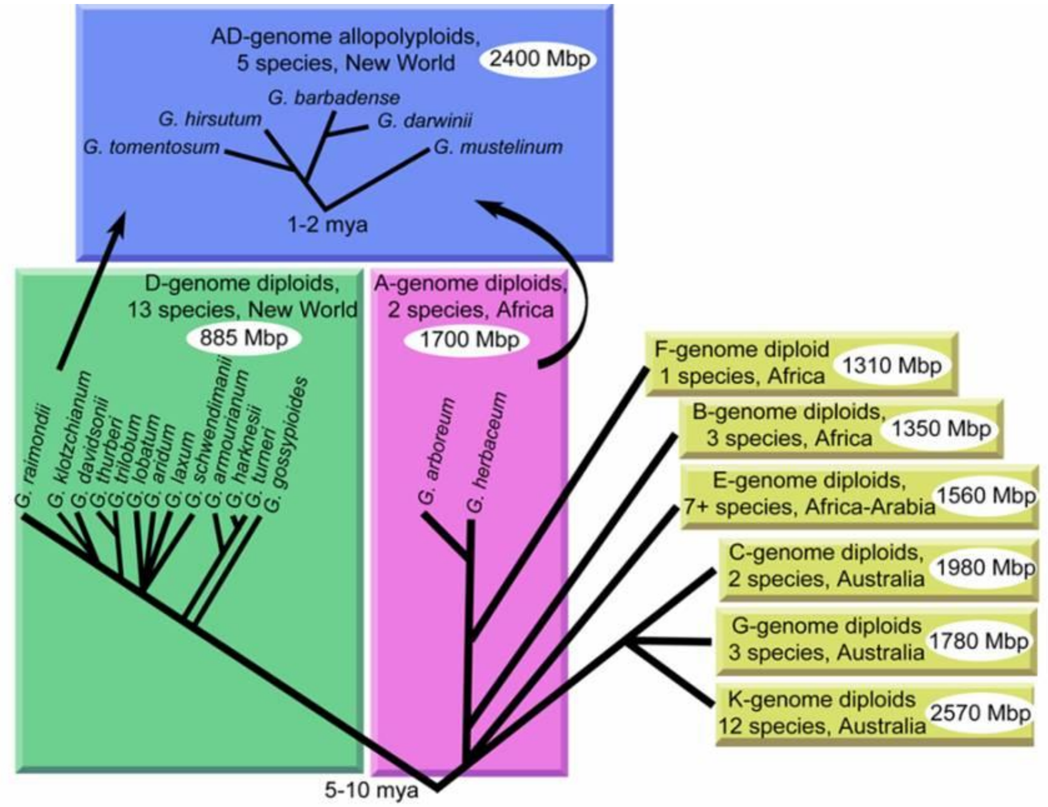


Figure 1.1 Evolutionary history of the Gossypium genus as depicted in Wendel et al. (2013), inferred from multiple phylogenetic data sets. (Not including the additional recently determined two allotetraploid species.)

Estimations of genome size (C-value) for many cotton species using flow cytometry methods have shown that A-genomes have nearly doubled in relative size, at 1700Mb, from the 885Mb D-genome diploids following divergence (Hendrix and Stewart 2005). This has been shown to be due primarily to the expansion of “retro” elements in the A-genome lineage following speciation (Hawkins, Kim et al. 2006).

Tetraploid species have been estimated to have genome sizes very close to the summation of the A- and D-diploid ancestor genomes at 2,400 Mb.

As the relative amounts and developmental patterns of seed fibers are very different between A- and D-genome diploids that formed the tetraploid, it is expected that the understanding of the underlying differences between these genomes will help to elucidate the formation of superior fiber characteristics in the allotetraploids. In order to study these differences at the sequence level, catalogues homozygous differences between the diploid genomes or “homeo-SNPs” have been developed (Page, Gingle et al. 2013). Utilization of this data in the tetraploid however is difficult in that the A- and D-genome diploids available today are not the exact ancestors of the tetraploid and many millions of years of evolution of these lines will allow for sequence variation that will not be informative in the allotetraploid lineage.

Need for Sequence-based Tools

Tools for detecting sequence differences between samples are necessary for genome-wide studies such as was completed to determine the evolutionary history of the *Gossypium* genus. These tools are also important for developing markers that can be utilized to facilitate marker-based crop improvement. In the history of cotton research, a number of different marker types have been utilized, including amplified fragment length polymorphisms (AFLPs), restriction fragment length polymorphisms (RFLPs), simple sequence repeats (SSRs), and recently single nucleotide polymorphisms (SNPs). Moderate density genetic linkage maps containing typically up to a couple thousand markers have been generated using RFLPs (Reinisch, Dong et al. 1994; Shappley, Jenkins et al. 1998), AFLPs (Lacape, Nguyen et al. 2003), and SSRs (Guo, Cai et al. 2007; Yu, Kohel et al. 2012). In 2012, a high-density consensus map was completed to produce the highest density map for cotton to date containing around ten thousand markers derived from multiple SSR linkage mapping projects (Blenda, Fang et al. 2012). Linkage maps have also been utilized in cotton for quantitative trait loci (QTL) localization, primarily using SSRs (Shen, Guo et al. 2007; Zhang, Hu et al. 2009; Zhang, Zhang et al. 2012).

While SNPs are the most abundant marker type throughout a genome, their discovery and use in linkage mapping in cotton has been limited until recently (An, Saha et al. 2008; Van Deynze, Stoffel et al. 2009; Buriev, Saha et al. 2010). The discovery of SNPs, particularly in cultivated cotton, has been limited because of the complex genome structure of cotton. The recent polyploidization between recently diverged diploids cause large homeologous regions of subgenomes to have high levels of sequence similarity and as recent expansion of the repetitive elements causes additional nearly identical regions throughout the genome. These sequence redundancies at two of more sites create difficulties for localizing and direct comparison of short sequencing reads that have typically been used to develop SNPs (Wang, Zhang et al. 2012). The polyploid nature creates complexity for identifying true allelic-SNPs apart from homeo-SNPs that appear as allelic-SNPs because of co-localization of homeologous sequences from subgenomes that cause homozygous differences between subgenomes to appear as SNPs (Kaur, Francki et al. 2012). Additionally because of ancient paleopolyploidy in the cotton lineage, discrimination between other similar regions caused by paralogs and orthologs is also difficult (Rong, Bowers et al. 2005).

Diversity in Cultivated Cotton

The low level of genetic diversity among upland cotton cultivars is the cumulative effect of several aspects of their origin. Not only was the ancestral AD species the product of a recent polyploidization event, but other events including divergence of *G. hirsutum* from other tetraploid species, subsequent domestication, and continuous selection of superior cultivated types which created additional genetic bottlenecks that further reduced genetic diversity among *G. hirsutum* breeding germplasm, particularly among cultivated lines. Multiple studies using primarily SSRs have shown that diversity present in elite lines is very limited (Fang, Hinze et al. 2013). It has been shown that the current Upland cultivar lines descend from a very small set of only about a dozen introgression events which has led to very limited diversity (Richmond 1950; Van Esbroeck and Bowman 1998). Due to the lack of variation among elite cotton germplasm, genetic variation for important agricultural elements for improvement of yield, adaptation to changing atmospheric conditions, water-use efficiency, abiotic and biotic stress resistance, and for minimizing fertilizer and pesticide requirements is unlikely to be sufficient within the currently available elite germplasm. Major genetic improvements are needed for cotton cultivars and production methods, and the genetic improvements will require more extensive use of Gossypium germplasm resources, as well as mutagenesis and various types of genetic engineering.

Importance of Uncultivated Cotton Species

While elite cultivated varieties exhibit only minimal amounts of variation, wild species harbor comparatively enormous levels of variation as well as unique genes (Lubbers, Chee et al. 2003). These beneficial variations when introgressed to elite germplasm can provide novel diversity for genetic improvement that is not possible within the currently available elite germplasm. Direct crosses between tetraploid species are relatively straightforward and have been utilized routinely. While direct crosses introgression from triploid F_1 hybrids between diploids and tetraploids is typically not practical due to the difficulty of obtained hybrids and sterility, a stable base chromosome number in the cotton genus, synthetic tetraploids and hexaploids with diploid derivatives have been created as bridges to be used for introgression of desired segments (Phillips and Strickland 1966; Brubaker, Brown et al. 1999; Robinson, Bell et al. 2007).

While introgression from wild species is possible, it is typically extremely time-consuming, labor-intensive and requires large amounts of funds. Due to linkage drag, beneficial segments that are introgressed are typically co-integrated with undesirable genes. So these linkages must be broken by nearby recombination events to achieve useable germplasm products. Thus far, the introgressions, recombination, and manipulations of these introgressed regions have been constrained by lack of high-throughput genome-wide markers and systems to use them cost effectively for diverse applications.

Towards Allotetraploid Genome Assembly

High-degrees of collinearity and high levels of conservation of gene orders have been observed in the cotton genus (Rong, Abbey et al. 2004; Rong, Bowers et al. 2005). The genomes of D-genome diploid species are smaller and less complex than cultivated cotton. This is especially true for *G. raimondii* (D_5), which is putatively the most closely related extant D-genome species (Chen, Scheffler et al. 2007). Therefore it was targeted as the initial cotton genome to be sequenced. Two efforts that produced draft genome sequences for D_5 were published at nearly the same time (Paterson, Wendel et al. 2012; Wang, Wang et al. 2012), however the Joint Genome Institute (JGI) reference genome sequence produced using a BAC-by-BAC approach is regarded by the cotton community as the superior reference (Paterson, Wendel et al. 2012). Intuitively the proposed second cotton genome to be sequenced was the A_2-genome diploid *G. arboreum*, an extant relative to the A-subgenome donor to the allotetraploid. A draft reference sequence for A_2 was recently published, but it is highly fragmented (Li, Fan et al. 2014).The availability of genome sequence assemblies from A- and D- genomes related to ancestors of the polyploidization event was expected by many to directly impact practical uses with the AD genome of cultivated cotton and other tetraploid relatives.

However direct utilization of the diploid genome sequences for use in sequencing the tetraploid is still extremely complicated and largely compromised by genomic redundancies, polyploid nature, recently amplified repetitive DNA families and numerous other complexities found in the evolutionarily diverged AD genome of *G. hirsutum*. Thus multiple tools need to be developed to assist and develop a high-quality draft genome for cultivated cotton. The proposed strategy for development of a BAC-based AD draft genome sequence includes the establishment of a physical map and minimum tiling path of finger-printed contigs of *G. hirsutum* homeologous chromosomes (Chen, Scheffler et al. 2007). This goal can be facilitated by development and integration of anchored DNA markers in linkage maps and BAC-end sequences that can be validated with alternative methods such as radiation hybrid mapping and/or fluorescent in situ hybridization of BACs (Wang, Song et al. 2006; Chen, Scheffler et al. 2007). High-density linkage and/or physical maps with (10,000+ markers) are typically used with sequence information to assist and validated sequence assemblies (Lewin, Larkin et al. 2009). The further development of large numbers of markers, particularly SNPs, will allow for development of comprehensive interspecific linkage maps as well as intraspecific linkage maps that can be used to anchor and assemble genomic sequences.

Completion of an accurate draft genome sequence for tetraploid cotton will expedite diverse areas of cotton research such as, traditional plant breeding, genomic selection, marker-assisted selection, gene dissection and biochemical pathway analysis of fiber initiation and development. It will also stimulate fundamental research on genome evolution during and following polyploidization, gene expression, cell differentiation and development, and cellulose synthesis. Many additional opportunities and advances will result from completion of a draft genome sequence for cultivated cotton. Some will be profound, such as enhanced rate and range of genetic manipulations. Ramifications will be extensive, leading to increase in yield and quality, germplasm utilization, and sustainability, eg. heat and drought tolerance. As shown in other crops (ie. maize, rice, sorghum) upon release of their genome sequences, contemporary genomic technologies will be rapidly assimilated into cotton research and genetic improvement. These kinds of improvements are needed by society if it is to provide for itself in a sustainable manner, simultaneously contending with population growth, decreased farming acreage, availability and quality of fresh water resources and global warming.

In order to assist in assembling the cultivated allotetraploid *Gossypium hirsutum* L. draft genome sequence, I have [1] developed and mapped gene-associated interspecific SNPs for five species, *G. barbadense*, *G. tomentosum*, *G. mustelinum*, *G. armourianum*, and *G. longicalyx*, [2] derived BAC-associated intraspecific and interspecific SNPs utilizing resequencing data, [3] produced and validated the first standardized high-throughput genotyping array for cotton, the CottonSNP63K, and [4] developed methods for localization of sequences to allotetraploid cotton chromosomes using resequencing of interspecific F_1 hypo-aneuploids. These various efforts provide additional validation and placement methods that can be directly integrated with the physical map constructed for cultivated cotton in order to ultimately produce a high-quality draft genome sequence.

CHAPTER II

DEVELOPMENT AND BIN MAPPING OF GENE-ASSOCIATED INTERSPECIFIC SNPS FOR COTTON (*GOSSYPIUM HIRSUTUM* L.) INTROGRESSION BREEDING EFFORTS[*]

Introduction

Cotton (*Gossypium spp.*) is the leading natural fiber crop worldwide and an important contributor to the economies of nearly 100 countries. The genus *Gossypium* is also an important model species for polyploidy and the biological processes of cell wall elongation and cellulose biosynthesis in fiber cells. This clade consists of approximately 45 diploid species and five allotetraploid species. Genomes of the allotetraploid species have 52 chromosomes (2n = 4x = 52) and are believed to have originated from a single polyploidization event between an A-genome diploid (n = 2x = 26) and a D-genome diploid (n = 2x = 26) approximately 1–2 million years ago (Wendel, Brubaker et al. 2009). The five allotetraploid species share a basic AD genome architecture. Chromosomes of the *G. hirsutum* genome ($[AD]_1$) have been numbered according to their evolutionary origins and meiotic pairing relationships. Chromosomes 1–13 comprise the "A" sub-genome (A_T) that originated from the extinct A-genome diploid ancestor and chromosomes 14–26 comprise the "D" sub-genome (D_T) that originated

[*] Reprinted with permission from "Development and bin mapping of gene-associated interspecific SNPs in cotton (*Gossypium hirsutum* L.) introgression breeding efforts" by AM Hulse-Kemp, H Ashrafi, X Zheng, F Wang, KA Hoegenauer, ABV Maeda, SS Yang, K Stoffel, M Matvienko, K Clemons, JA Udall, A Van Deynze, DC Jones, and DM Stelly, 2014. *BMC Genomics*. **15**:945, Copyright 2014 by AM Hulse-Kemp.

from the extinct D-genome diploid ancestor. There are four major cultivated species wordwide, two diploids *G. arboreum* (A_2 genome) and *G. herbaceum* (A_1) and two allotetraploids *G. hirsutum* L. or Upland cotton and *G. barbadense* L. ($[AD]_2$), extra long-staple Pima, Egyptian cotton or Sea Island cotton. Upland cotton cultivation represents over 95% of the fiber produced worldwide due to its high yield, but generally Pima cotton the next most cultivated cotton, exhibits longer, stronger, and finer fiber.

Upland cotton has a very narrow genetic base due to multiple bottleneck events including, polyploidization, domestication and continuous selection. It has been suggested and experimentally tested that current Upland cultivars descend from only about a dozen introgressions and therefore exhibit an extremely small amount of diversity (Richmond 1950; Van Esbroeck and Bowman 1998). With such small diversity in elite cotton germplasm, it is unlikely that sufficient variation for agronomically important traits, such as, fiber properties, yield, disease and insect resistance, drought tolerance and changing atmospheric conditions will be found within currently available elite breeding germplasm. Wild cotton species harbor large numbers of unique genes, which upon introgression may provide novel diversity for genetic improvement.

Diploid cotton species have been shown to have many disease and insect resistance traits, as well as improved fiber characteristics. The diploid *G. longicalyx* Hutch and Lee (F_1) is the only member of the F-genome clade and is native to Africa. It has been shown to have resistance to pathogens, such as reniform nematode (Yik and Birchfield 1984), and to have beneficial genes for fiber quality (Weaver, Sikkens et al. 2013). The diploid *G. armourianum* Kearney (D_{2-1}) belongs to the D-genome clade and is a wild species found in Mexico. It has been shown to exhibit resistance to the whitefly (Jayaraj and Palaniswamy 2005), which is the vector for many cotton pathogens such as the leaf curl virus (Briddon and Markham 2000). The diploid species exhibit a large range of relative genome sizes. Due to the difference in chromosome number between diploids and cultivated cotton, methods to move genes from diploids into cultivated tetraploid cotton using synthetic tri-species hybrids have been devised to introgress desired diploid segments through breeding (Mergeai 2006; Robinson, Bell et al. 2007).

While crossing cultivated tetraploid cotton directly with diploid species is difficult, allotetraploid species can be easily interbred and then backcrossed to move desired segments into cultivated material. The tetraploid *G. tomentosum* Nuttal ex Seeman originates from the Hawaiian islands and produces a small amount of short, reddish brown fiber. *G. tomentosum* has been found to show resistance against the cotton leaf hopper, *Amrasca biguttula biguttula*, and thrips, *Frankliniella occidentalis* (Jayaraj and Palaniswamy 2005). The tetraploid *G. mustelinum* Meers ex Watt is from Brazil and also produces a small amount of lint. Using HPLC analysis, *G. mustelinum* has been shown to have the highest leaf concentrations of terpenoid aldehydes that affect insect resistance (Altaf, Stewart et al. 1997). *G. barbadense* originates from South America and is a cultivated species which represents about five percent of the annual worldwide fiber crop. This tetraploid exhibits excellent fiber quality characteristics for fiber length, micronaire and high strength relative to *G. hirsutum*.

Many of the mapping efforts in cotton have consisted of interspecific biparental populations of *G. hirsutum* × *G. barbadense* which offers a higher polymorphism rate than intraspecific crosses, and segregation for superior fiber quality characteristics. Moderate density linkage maps have been created using restriction fragment length polymorphisms (RFLPs) (Reinisch, Dong et al. 1994; Shappley, Jenkins et al. 1998), amplified fragment length polymorphisms (AFLPs) (Lacape, Nguyen et al. 2003) and simple sequence repeats (SSRs) (Guo, Cai et al. 2007; Yu, Kohel et al. 2012). SSRs have also been used for wide-cross whole-genome radiation hybrid (WWRH) mapping for production of syntenic groups (Gao, Chen et al. 2004; Gao, Chen et al. 2006). A consensus map was recently created which integrated all of the previous mapping efforts (Blenda, Fang et al. 2012). While single nucleotide polymorphisms (SNPs) represent the most prevalent category of polymorphisms available within the genome, few studies have developed and mapped SNPs in cotton (Byers, Harker et al. 2012; Yu, Kohel et al. 2012). SNP development efforts to-date have produced relatively few numbers of SNPs using different genome reduction methods in cultivated species (An, Saha et al. 2008; Van Deynze, Stoffel et al. 2009; Buriev, Saha et al. 2010; Byers, Harker et al. 2012; Rai, Singh et al. 2013).

An aspect of polyploid genomes that creates difficulties during SNP development is that there are two indistinguishable types of SNPs in polyploid sequence data: homeologous sequence variants or “homeo-SNPs” and traditional SNPs or “allele-SNPs” (Kaur, Francki et al. 2012). A catalogue of homeo-SNPs, which are differences between the A-genome and D-genome diploid species, was recently identified in *Gossypium* diploid and tetraploid genomes (Page, Gingle et al. 2013; Page, Huynh et al. 2013). In cotton tetraploids, five million homeo-SNPs were found between the A_T and D_T

subgenomes, which was facilitated by recent publication of the reference genome sequence for *G. raimondii* (D_5) (Paterson, Wendel et al. 2012). The D_5 genome is regarded as the closest living diploid relative to the D-genome ancestor of current AD-allotetraploid species (Wendel and Cronn 2003). It has been hypothesized that the catalogued homeo-SNPs can possibly be used to filter putative SNPs when sequence reads are aligned within the framework of the base-pair coordinates of reference diploid genomes (Paterson, Wendel et al. 2012; Li, Fan et al. 2014). While homeo-SNPs may allow for separation of homeologous sequences, they are not directly applicable to breeding. As allele-SNPs identify polymorphisms within a haplotype, experimental assays can be developed to genotype individuals and track favorable and unfavorable alleles. Upon germplasm introgression from wild species, whether diploid or tetraploid, orthologous sequence variants become allele-SNPs. High-density interspecific allele-SNPs distributed across both sets of *G. hirsutum* chromosomes will be useful for breeders to efficiently introgress quantitative trait loci (QTLs) and track alleles in marker-assisted selection (MAS) of beneficial traits from donor species.

Traditionally, interspecific introgression breeding efforts are extremely time-consuming and require large amounts of effort and funds. Interspecific genetic introgression into *G. hirsutum* has thus far been constrained by the paucity of high-throughput genome-wide markers that would facilitate tracking of introgressed segments. The relative scarcity of SNPs in cultivated allotetraploid cotton reflects the difficulty of developing SNPs for its complex genome, comprised of large repetitive regions and homeologous content due to recent polyploidization. Here, we report a method utilizing the genomic reduction method of transcriptome sequencing to derive interspecific gene-associated SNPs between the genetic standard *G. hirsutum* TM-1, and five other species, including the genetic standard *G. barbadense* doubled haploid line 3–79, two allotetraploids *G. tomentosum* and *G. mustelinum,* and two diploids *G. armourianum* and *G. longicalyx.* These SNPs will be extremely beneficial for high-density interspecific mapping and will help revolutionize introgression breeding efforts by facilitating MAS-based introgression, genetic dissection, and gene utilization in cultivated cotton.

Materials and Methods

Plant materials

The seed of *G. barbadense* L. $(AD)_2$ genetic standard line 3–79, *G. tomentosum* $(AD)_3$ plant number 19909036.05 from the Beasley Lab collection, *G. mustelinum* $(AD)_4$ plant number 200508123.02 from the Beasley Lab collection, *G. armourianum* (D_{2-1}) accession D2-1-6, and *G. longicalyx* (F_1) plant number 200908137.04 from the Beasley Lab collection were planted at Texas A&M University. Young leaf tissues were sampled from each plant and used to isolate total RNA using the Qiagen RNeasy Mini Kit per manufacturer instructions. RNA isolates were quantified using NanoDrop spectrophotometry (Thermo Scientific, Wilmington, USA) and checked for quality by gel electrophoresis. PolyA RNA was extracted using double purification with oligo dT Dynal beads. Illumina RNA-Seq libraries were prepared using the manufacturer protocol (Illumina Inc, San Diego, USA). Libraries were normalized by denaturation and rehybridization in NaCl and TMAC (tetra-methyl-ammonium-chloride) buffers (Matvienko, Kozik et al. 2013) and then treated with Duplex Specific Nuclease to digest cDNAs from highly abundant transcripts (Zhulidov, Bogdanova et al. 2004). The treated library was then re-amplified for 12 PCR cycles using Illumina library primers. The libraries were then single-read sequenced using the Illumina Genome Analyzer II for 85 cycles. Raw single-read sequence files were uploaded to NCBI under BioProject PRJNA203021 and SRA numbers (SRX457172 – *G. barbadense*, SRX472724 - *G. tomentosum*, SRX - 474879/SRR174699 *G. mustelinum*, SRX474240/SRR1174039 and SRR1174041- *G. armourianum*, and SRX474242/SRR1174179 and SRR1174182 - *G. longicalyx*).

Wide-cross whole-genome radiation (WWRH) individuals (Gao, Chen et al. 2004) were planted and maintained at Texas A&M University. Small leaves were collected from each plant and extracted using the Qiagen DNeasy plant extraction kit per manufacturer instructions.

SNP development

Reads from each species were trimmed for quality and then aligned to the *G. hirsutum* L. assembly created from genetic standard line, TM-1 (GALV00000000.1 Ashrafi et al. - in preparation), using CLC Genomics Workbench (V5.0). Reads which mapped to multiple locations were randomly assigned to a single location. Putative SNPs between TM-1 and each species were identified one accession at a time. The mapping data were exported as *BAM* files to a Linux server and SAMtools were used to call variants (Li, Handsaker et al. 2009). Subsequent rounds of parameter tweaking resulted in the final pipeline used for SNP development. The resulting pileup files were filtered using the filter pileup Perl script in Galaxy (https://main.g2.bx.psu.edu/) (Goecks, Nekrutenko et al. 2010) to remove indels and positions with less than coverage of 3. The resulting file was then further filtered using an in-house Perl script which required two genotypes to be homozygous for different bases with minimum coverage of 10. Putative SNPs were then removed from the list if they were located within 50 bases of predicted intron-exon boundary on the TM-1 assembly using SGN (http://solgenomics.net/) intron finder tool. SNPs were further filtered to remove theoretical homeo-SNP positions based on allele-SNP calls generated when Illumina TM-1 reads were mapped back to the TM-1 reference.

SNPs from each species were classified based on identification of additional SNPs. Class I was defined as SNPs from contigs with no other SNP residing within the contig. Class II was defined as SNPs from contigs that contained one or more additional SNP outside of the 50-bp flanking sequences (none within). Class III was defined as SNPs from contigs that contained one or more additional SNPs within the 50-bp flanking sequences.

Removal of redundant markers

A FASTA file containing all *in silico*-derived SNPs was used for BLAST analysis (v2.2.27) against a FASTA file containing all *G. hirsutum* markers from the Ashrafi et al. (in preparation) dataset. Markers were removed from the *in silico* set if

BLAST analysis showed 100% identity over 100% length of the sequence to the *G. hirsutum* data set. The remaining set was then BLASTed against itself to determine identical markers within the set. Markers with hits in other species groups were removed in a hierarchical method, leaving markers in the highest set, based on the hierarchy *G. barbadense, G. tomentosum, G. mustelinum, G. armourianum,* and *G. longicalyx.* Markers with hits within the same species were separated into a data set containing "overlap markers". The result was a final non-redundant data set of Class I and Class II markers for each species. The final non-redundant set of Class I and Class II SNPs for each species were submitted to NCBI's dbSNP (ss974702651-ss974710721 for *G. barbadense,* ss1026506434-ss1026511516 for *G. tomensotum,* ss1026511517-ss1026516811 for *G. mustelinum,* ss1026516812-ss1026536246 for *G. armourianum,* and ss1026536247-ss1026565496 for *G. longicalyx*). Non-redundant Class III SNPs were compiled a file for all species.

Alignment of markers to D_5-reference genome

The non-redundant SNPs and overlap markers were aligned to the *G. raimondii* (D_5) reference genome sequence (Paterson, Wendel et al. 2012) using Burrows-Wheeler Aligner (BWA) in Galaxy (Goecks, Nekrutenko et al. 2010) using default settings. Alignment positions were corrected to note the SNP base positions. SNPs were separated into files based on D_5 genome scaffold alignment. Scaffold files were sorted by genome position. Densities of markers along the D_5 genome scaffolds were plotted using densityPlot in the R program (R Development Core Team 2010) over scaffold position.

SNP validation

Subsets of SNPs from the subsequent rounds of bioinformatic filtering were selected for experimental testing using the LGC KASP assays (Beverley, USA). Assay primers were developed using BatchPrimer3 with an optimal primer Tm of 57°C (minimum 55°C, maximum 60°C, maximum difference between primers of 5°C), optimal product size of 50 base pairs (minimum 50 base pairs, maximum 100 base pairs) and the default settings were used for the remaining parameters. Primers were mixed at the dilutions specified by LGC then used to perform KASP assays on small screening panels containing duplicates, e.g., the panel used to screen *G. barbadense* markers contained 2 *G. hirsutum L.* TM-1, 2 *G. barbadense* 3–79, 2 euploid F1 (*G. hirsutum x G. barbadense*), 4 RILs from a *G. hirsutum* × *G. barbadense* mapping population, and 2 non-template (negative) controls. Plates were initially run for the recommended 38 cycles on the LGC SNP platform, centrifuged then read on the Pherastar plate reader. The Pherastar files were imported into KlusterCaller software for genotyping. If the plates were determined to be insufficiently clustered, additional sets of 3 cycles were added and the plates were re-read and re-imported, until scoreable clusters were formed or the marker was deemed to be unacceptable.

Wide-cross whole-genome radiation hybrid mapping

Those *G. barbadense* markers which clustered well from all rounds of development used for parameter tweaking until the final pipeline was reached were then run on a “full-panel” containing duplicates of the parental and F_1 controls (*G. hirsutum* line TM-1, *G. barbadense* cultivar 3–79, euploid F1), *G. hirsutum* cultivar DP-90, *G. barbadense* cultivar Phytogen800, 131 wide-cross whole-genome radiation hybrid individuals, 47 F_1 hypo-aneuploid lines, and 4 non-template negative controls. Genotypes were manually curated. All questionable genotypes were listed as unscored. A shift of genotype from the F1 heterozygote cluster to the homozygous 3–79 cluster was interpreted as a deletion in the respective interspecific WWRH or hypo-aneuploid F_1 cytogenetic stock. Genotype files were manipulated to note presence (1) or absence (0) of a deletion in the WWRH plants.

An additional set of markers from two previous studies, Van Deynze et al. (2009) and Byers et al. (2012), were also genotyped using KASP assays. Genotypes for these markers were also manipulated to note presence or absence of a deletion in the WWRH plants.

SNP markers which showed no deletions among the 131 WWRH samples were removed from the WWRH mapping analysis. Genotype files in binary format were analyzed using Carthagene, with a LOD score of 3.0 and mapping distance within 100 cR. From the resulting syntenic groups, the singleton groups were removed and classified as non-linked markers. Those syntenic groups with two or more markers were subjected to finishing methods using annealing, flips and polishing to determine the final group order. The resulting syntenic groups were cross-referenced with the D_5 alignments of individual markers, as well as the chromosome locations determined by deletion analysis with the F_1 hypo-aneuploids. Syntenic groups and their relationships based on alignment to the D_5 scaffolds were plotted using Strudel software (Bayer, Milne et al. 2011).

Deletions analysis

The numerical distribution of deletions in the radiation hybrids per marker was analyzed by a box plot method to determine the number of deletions that would be statistically significant. The first and third quartile were determined along with the mean, then the inner quartile range (IQR) was utilized to calculate the threshold for outliers greater than 1.5 times the IQR. Markers which had numbers of deletions beyond the threshold were determined to have a significantly different number of deletions than expected.

Functional analysis

Contigs for which the SNPs represented identical differences from *G. hirsutum* TM-1 for all five species from the reference were parsed into a FASTA file containing 117 contigs for 118 SNP. The reference FASTA file was modified to contain the alternate SNP allele(s) using an in house Perl script. Both the reference and alternate FASTA files were analyzed using AUGUSTUS (Keller, Kollmar et al. 2011) to predict translation start and end sites using *Theobroma cacao* as the model species. Non-synonymous and synonymous changes between the reference and alternate files were calculated. Predicted amino acid coding sequences were then investigated using TBLASTX against NCBI’s non-redundant database. BLAST results were parsed to contain the top hit with covcutoff of 50 and *e* value cut off of 1^{e-8}. The differences between predicted protein products were investigated. The same analysis was performed using a randomly selected set of 118 SNP from the final overall Class I and Class II data set.

Results

SNP development

Utilizing the *G. hirsutum* (line TM-1) transcriptome assembly produced by Ashrafi et al. (in preparation) consisting of 72,450 contigs covering over 70 M bp with N50 of 1,100 bp, transcriptome sequence reads (**Table 2.1**) were aligned and utilized to identity and filter a total of 10,888 SNPs *in silico* for *G. barbadense* line 3–79 relative to *G. hirsutum* line TM-1. With the same bioinformatic pipeline, SNPs were also developed for *G. tomentosum* (9,520), *G. mustelinum* (10,988), *G. armourianum* (26,974), and *G. longicalyx* (38,217). Reads which mapped to multiple locations were randomly assigned to a single location to achieve higher mapping coverage with limited number of reads. Filtering included removal of theoretical homeo-SNP positions based on an index created by mapping back Illumina TM-1 reads to the assembly. All of the markers identified within a given species were classified according to surrounding polymorphisms for the same species. Marker classifications were based only on species-specific polymorphism data and determined independently for each species. “Class I” was a SNP in which no additional polymorphism was found to exist in the same contig. “Class II” was a SNP in which (an) additional polymorphism(s) was found within the same contig, but the additional polymorphism was outside of 50 base pairs (bp) of the marker. “Class III” was a SNP in which additional polymorphisms were found in the same contig and within 50 bp of the marker. The SNPs for *G. barbadense*, *G. tomentosum*, *G. mustelinum*, *G. armourianum* and *G. longicalyx* were classified according to these criteria (**Table 2.2**).

Table 2.1 Transcriptome sequence information.

Species	Sample	Raw reads (#)	Trimmed reads (#)	Mapped reads (#)	Depth	Reference coverage (%)
G. barbadense	3-79	101,276,621	101,276,621	43,519,596	48.45	84%
G. tomentosum	19909036.05	63,119,599	63,118,203	38,439,408	31.77	87%
G. mustelinum	200508123.02	65,940,564	65,940,111	40,767,777	33.07	86%
G. armourianum	D2-1-6	53,279,426	53,279,087	20,266,019	20.90	68%
G. longicalyx	200908137.04	52,050,537	52,050,305	23,393,277	24.50	65%

* Raw and processed read information of Illumina GA-II (Illumina) sequence generated from RNA-Seq libraries for *G. barbadense, G. tomentosum*, *G. mustelinum*, *G. armourianum*, and *G. longicalyx*.

Table 2.2 List of unfiltered SNPs determined for all species.

	Class I	Class II	Class III	Total
G. barbadense	3,257	6,385	1,246	10,888
G. tomentosum	1,520	6,526	1,474	9,520
G. mustelinum	1,678	7,584	1,726	10,988
G. armourianum	7,331	14,523	5,120	26,974
G. longicalyx	14,546	18,960	4,711	38,217

Number of SNPs derived *in silico* relative to *G. hirsutum* inbred line TM-1 for species *G. barbadense*, *G. tomentosum*, *G. mustelinum*, *G. armourianum*, and *G. longicalyx*. SNPs are classified into three categories, Class I are SNPs from contigs with no other SNP residing within the contig. Class II are SNPs from contigs that contain one or more additional SNP outside of the 50-bp flanking sequences (none within). Class III are SNPs from contigs that contain one or more additional SNPs within the 50-bp flanking sequences.

Removal of redundant markers

SNPs that were redundant across species were reduced to a single instance by means of progressive comparisons (see Methods). The overlap with intraspecific *G. hirsutum* markers (Ashrafi et al. - in preparation) was low, e.g., only 3.3% or 367 markers of the *G. hirsutum-G. barbadense* SNPs were found to be redundant compared to the intraspecific SNPs. The overlap among the different species sets was plotted in a Venn Diagram, which revealed moderate levels of overlap among the three AD species (**Figure 2.1**). Following the stated progression, a total of 10,521 non-redundant *G. barbadense* SNPs were identified, 2,647 were Class I, 5,660 were Class II (**Additional file 2.1**), and 1,189 were Class III (**Additional file 2.2**). In addition, when the *G. barbadense* set was BLASTed against itself, 1,025 markers were redundant within the *G. barbadense* set (**Additional file 2.3**). SNPs of this redundant nature have been identified and listed separately for each species, so that they can be avoided (or targeted) for future species-specific studies on alternative splicing or gene family composition. For *G. tomentosum* 6,396 SNPs were retained, 811 in Class I, 3,885 in Class II (**Additional file 2.4**), and 1,107 in Class III (**Additional file 2.2**), while 593 redundant SNPs were listed separately (**Additional file 2.5**). A total of 6,663 SNPs were retained for *G. mustelinum*, 822 in Class I, 4,085 in Class II (**Additional file 2.6**), 1,107 in Class III (**Additional file 2.2**) and 592 redundant SNPs (**Additional file 2.7**). For *G. armourianum*, 5,723 Class I, 12,033 Class II (**Additional file 2.8**), 4,648 Class III (**Additional file 2.2**) and 2,425 redundant SNPs (**Additional file 2.9**) were obtained for a total of 24,829 SNPs. Lastly for *G. longicalyx* a total of 34,550 SNPs were identified, including 11,435 in Class I, 15,454 in Class II (**Additional file 2.10**), 4,309 in Class III (**Additional file 2.2**) and 3,352 SNPs which were redundant and listed separately (**Additional file 2.11**). In the final set of 62,555 non-redundant Class I and Class II SNPs for all of the five species, the transition to transversion ratio was 1.63 (38,763/23,792). Non-redundant SNPs were unique in the final set for the SNP and 50bp flanking sequences (now reference as SNP markers).

Alignment of markers to D_5-reference genome

A moderate share (75.87%) or 47,672 SNP markers in the final set could be aligned to the *Gossypium raimondii* (D_5) diploid reference genome sequence. *G. armourianum* had the highest percentage of mapped markers, followed by *G. longicalyx,* and the three tetraploids *G. barbadense, G. tomentosum,* and *G. mustelinum*. Nearly all of the mapped markers (99.7%) aligned to one of the thirteen pseudo-chromosome scaffolds, and only 154 (0.3%) markers were aligned to unplaced scaffolds. A bimodal distribution across each D_5-chromosome was observed when average density of markers was plotted (**Figure 2.2**).

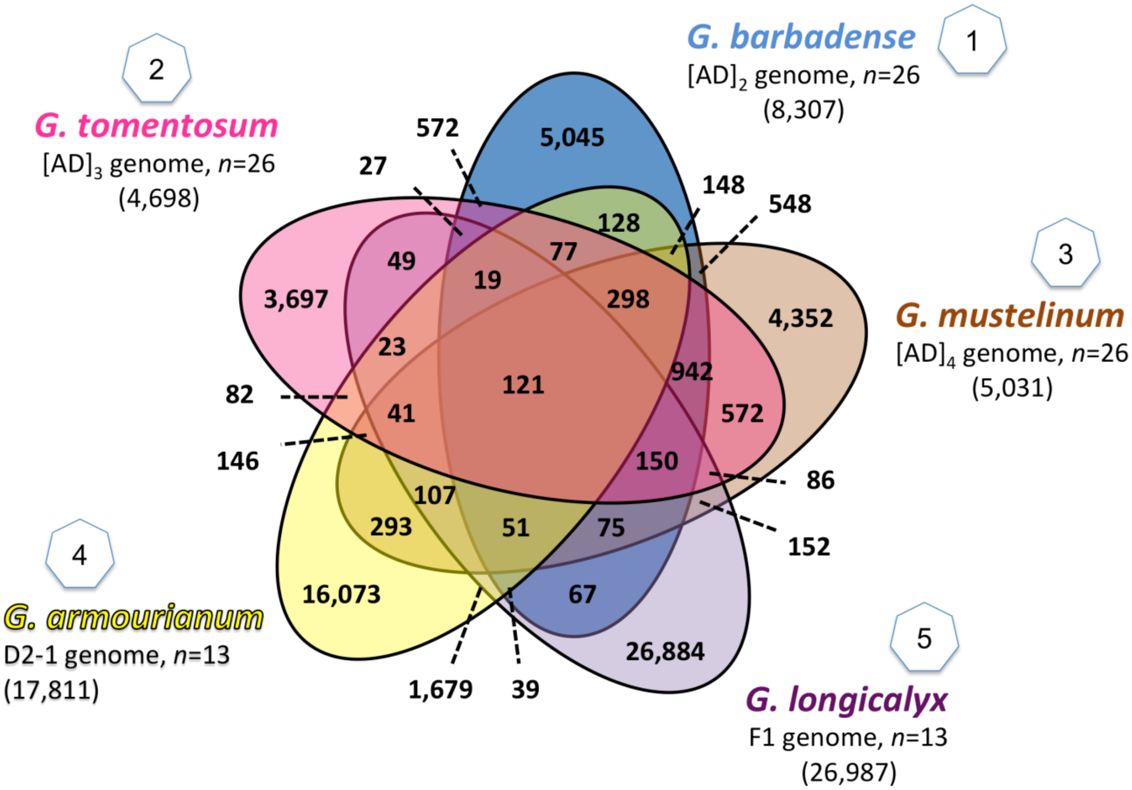


Figure 2.1 Overlap of SNPs among species. The overlap and specificity of the Class I and Class II SNPs for *G. barbadense* cv. 3–79, *G. tomentosum*, *G. mustelinum*, *G. armourianum*, and *G. longicalyx.*

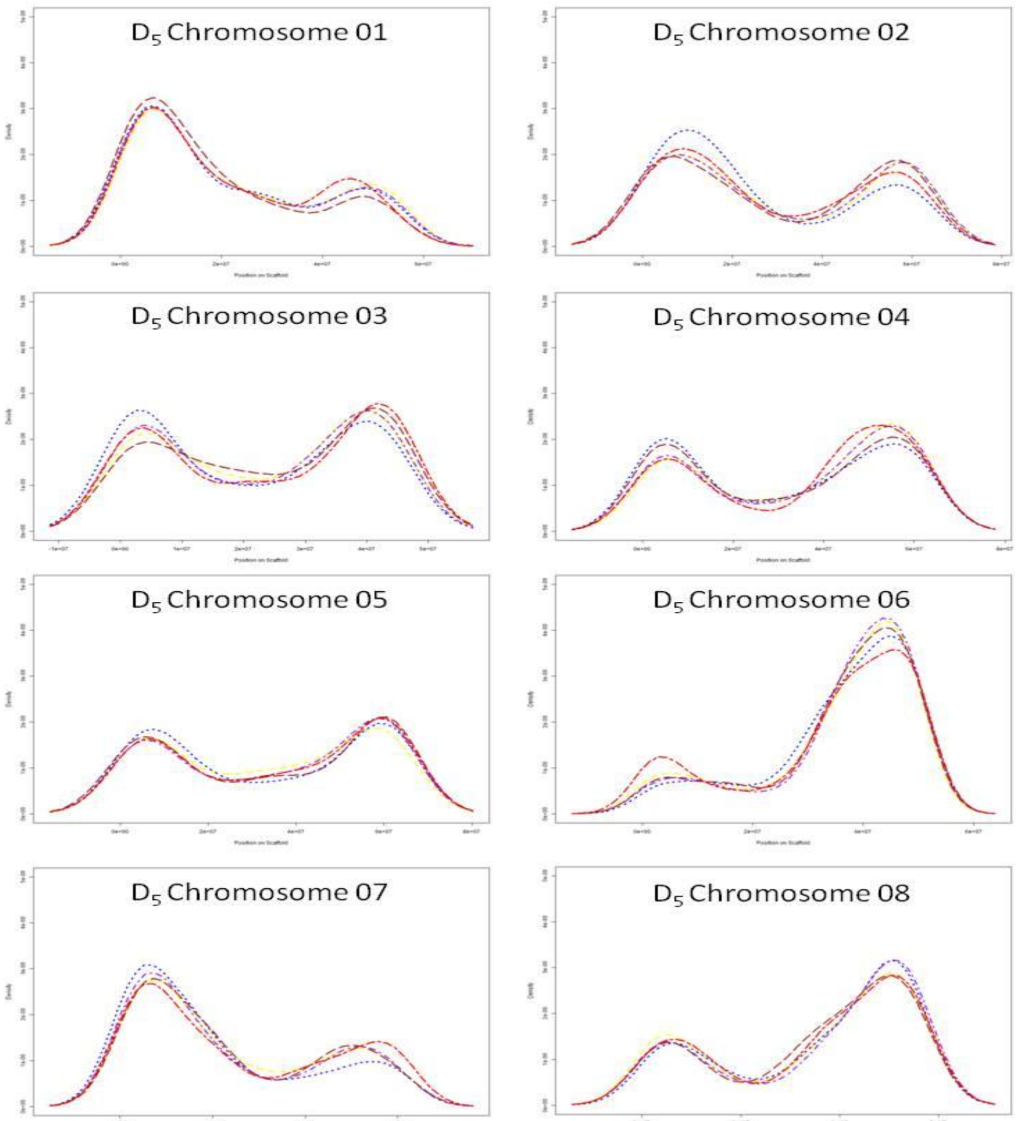


Figure 2.2 Distributions of SNPs relative to *Gossypium raimondii* (D_5) draft genome. All SNPs for each species were plotted according to BWA alignment positions (X-axis) across the *G. raimondii* (D_5) draft genome over a sliding window in R. Density (Y-axis) is the proportion of the number of SNPs within a species-specific data set calculated over a sliding window. A figure was produced for each of the 13 scaffolds of the draft genome sequence. *G. barbadense* cv. 3–79 is shown in blue, *G. tomentosum* is shown in red, *G. mustelinum* is shown in brown, *G. armourianum* is shown in yellow, *G. longicalyx* is shown in purple.

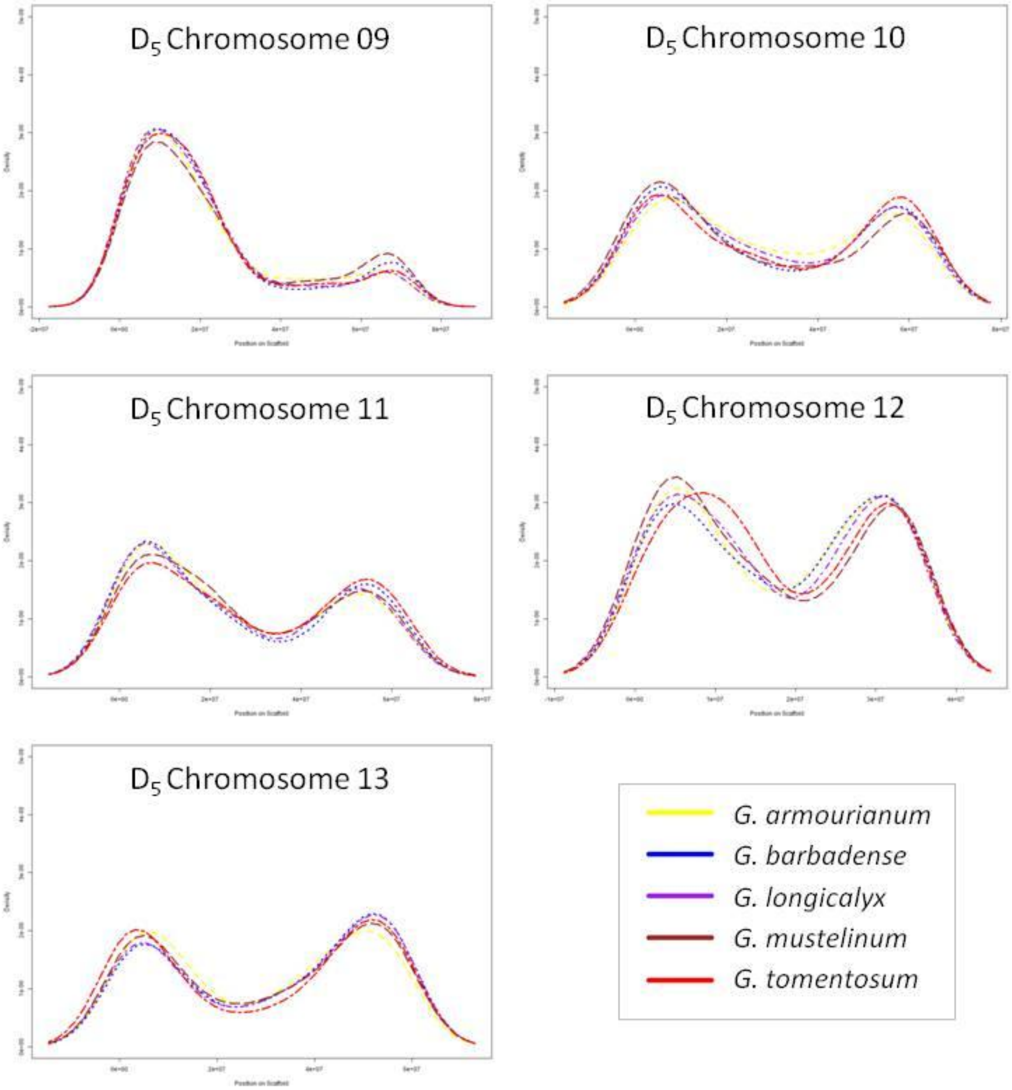


Figure 2.2 (continued)

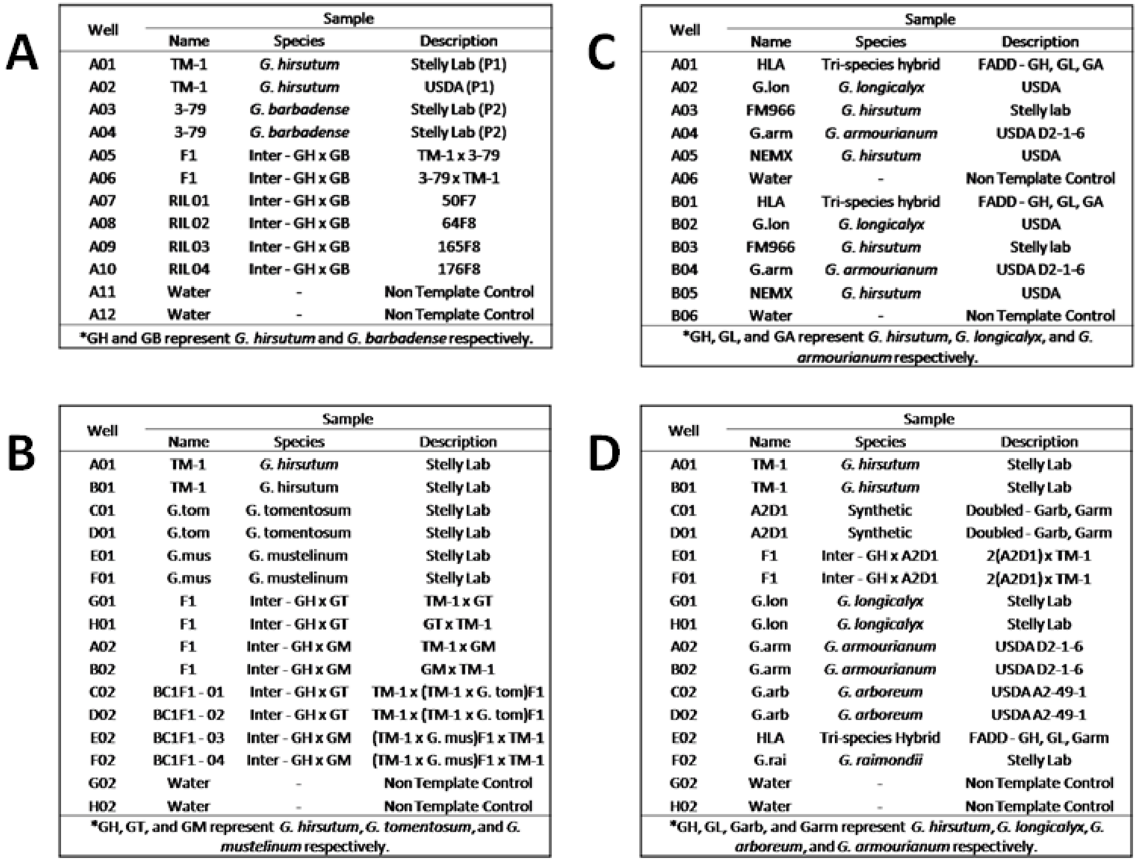
Validation of SNPs

Random sets of markers from the non-redundant final set of Class I and Class II SNPs were tested using KASP end-point assays (LGC Genomics, Beverly, MA, USA) from each species. As random sets of markers for each species were selected for screening prior to development of the final non-redundant set, some markers may have been tested on species other than the ones with which they were associated in the non-redundant list of SNPs (**Additional files 2.2, 2.3, 2.4, 2.5, 2.6, 2.7, 2.8, 2.9, 2.10 and 2.11**). This is due to the fact that some markers detected the same SNP between *G. hirsutum* and multiple species and was retained only once in the non-redundant list. The validation status of each marker for tested species is noted in columns B and C of **Additional files 2.2, 2.3, 2.4, 2.5, 2.6, 2.7, 2.8, 2.9, 2.10 and 2.11**.

In the final non-redundant data sets, a total of 665 randomly selected markers were tested from the *G. barbadense* Class I and Class II set. Of these, 262 markers were tested on the "*G. barbadense* screening panel" (**Figure 2.3**) and 209 (79.8%) had clean clusters which allowed for scoring P1, P2 and F1 genotypes (successful markers). The remaining 403 markers were screened on panels from the other species (*G. tomentosum, G. mustelinum, G. armourianum* or *G. longicalyx*) and 286 (71%) of those were validated to have scoreable genotypes. Sets of markers were also screened which fall under the species-specific data sets, 466 were tested from the *G. tomentosum* set of which 252 were tested on the *G. tomentosum* screening panel which produced 168 (66.7%) successful markers. The other 214 markers were tested on other species and 141 (65.9%) were validated. A total of 138 were tested from the *G. mustelinum* data set, of which 90 were run on the *G. mustelinum* screening panel and 48 were run on other species screening panels. These tests resulted in 61 (67.8%) successful markers from the same-species tests and 27 (56.3%) successful markers from the other-species tests. For the *G. armourianum* data set, 214 markers were tested, of which only 5 were tested on species-specific panels. Overall, 146 out of the 209 (69.9%) markers tested on *G. armourianum* produced successful assays. A small set of 27 markers was tested from the *G. longicalyx* data set, of which 19 (70.4%) generated successful assays. A similar proportion of successful assays was obtained from SNPs generated for *G. longicalyx* using a different *G. hirsutum* assembly version that was abandoned because it yielded poor results when used to define SNPs in other species. Unique SNPs from this set were extracted and included in **Additional file 2.12**.

Validation of G. barbadense SNPs – Wide-cross whole-genome radiation hybrid mapping

A total of 509 markers were WWRH mapped using 131 WWRH individuals (*G. hirsutum* line TM-1 × irradiated pollen of *G. barbadense* cv. 3–79 (Gao, Chen et al. 2004)). Those markers (124) which were found using a previous version of the bioinformatic pipeline (which produced sets with overall lower success rate) and thus are not in the final set (**Additional files 2.2 and 2.3**) have names and sequences listed in **Additional file 2.13**. A total of 60 syntenic groups were produced along with 43 singletons (**Additional file 2.14**) that were not integrated into a syntenic group. Most of the groups (52) were anchored onto the *G. raimondii* draft genome sequence (**Figure 2.4**) by alignment of markers, as well as by chromosome localization using deficiency mapping with F_1 hypo-aneuploids (**Figure 2.5**) and/or the presence of marker(s) that were previously linkage-mapped (Yu, Kohel et al. 2012). The markers in **Figure 2.4** are reported in bins because the order of the markers may not be accurately estimated as the WWRH panel did not provide enough power to precisely order markers.



A

Well	Sample		
	Name	Species	Description
A01	TM-1	*G. hirsutum*	Stelly Lab (P1)
A02	TM-1	*G. hirsutum*	USDA (P1)
A03	3-79	*G. barbadense*	Stelly Lab (P2)
A04	3-79	*G. barbadense*	Stelly Lab (P2)
A05	F1	Inter - GH x GB	TM-1 x 3-79
A06	F1	Inter - GH x GB	3-79 x TM-1
A07	RIL 01	Inter - GH x GB	50F7
A08	RIL 02	Inter - GH x GB	64F8
A09	RIL 03	Inter - GH x GB	165F8
A10	RIL 04	Inter - GH x GB	176F8
A11	Water	-	Non Template Control
A12	Water	-	Non Template Control

*GH and GB represent *G. hirsutum* and *G. barbadense* respectively.

B

Well	Sample		
	Name	Species	Description
A01	TM-1	*G. hirsutum*	Stelly Lab
B01	TM-1	G. hirsutum	Stelly Lab
C01	G.tom	G. tomentosum	Stelly Lab
D01	G.tom	G. tomentosum	Stelly Lab
E01	G.mus	G. mustelinum	Stelly Lab
F01	G.mus	G. mustelinum	Stelly Lab
G01	F1	Inter - GH x GT	TM-1 x GT
H01	F1	Inter - GH x GT	GT x TM-1
A02	F1	Inter - GH x GM	TM-1 x GM
B02	F1	Inter - GH x GM	GM x TM-1
C02	BC1F1 - 01	Inter - GH x GT	TM-1 x (TM-1 x G. tom)F1
D02	BC1F1 - 02	Inter - GH x GT	TM-1 x (TM-1 x G. tom)F1
E02	BC1F1 - 03	Inter - GH x GM	(TM-1 x G. mus)F1 x TM-1
F02	BC1F1 - 04	Inter - GH x GM	(TM-1 x G. mus)F1 x TM-1
G02	Water	-	Non Template Control
H02	Water	-	Non Template Control

*GH, GT, and GM represent *G. hirsutum*, *G. tomentosum*, and *G. mustelinum* respectively.

C

Well	Sample		
	Name	Species	Description
A01	HLA	Tri-species hybrid	FADD - GH, GL, GA
A02	G.lon	*G. longicalyx*	USDA
A03	FM966	*G. hirsutum*	Stelly lab
A04	G.arm	*G. armourianum*	USDA D2-1-6
A05	NEMX	*G. hirsutum*	USDA
A06	Water	-	Non Template Control
B01	HLA	Tri-species hybrid	FADD - GH, GL, GA
B02	G.lon	*G. longicalyx*	USDA
B03	FM966	*G. hirsutum*	Stelly lab
B04	G.arm	*G. armourianum*	USDA D2-1-6
B05	NEMX	*G. hirsutum*	USDA
B06	Water	-	Non Template Control

*GH, GL, and GA represent *G. hirsutum*, *G. longicalyx*, and *G. armourianum* respectively.

D

Well	Sample		
	Name	Species	Description
A01	TM-1	*G. hirsutum*	Stelly Lab
B01	TM-1	*G. hirsutum*	Stelly Lab
C01	A2D1	Synthetic	Doubled - Garb, Garm
D01	A2D1	Synthetic	Doubled - Garb, Garm
E01	F1	Inter - GH x A2D1	2(A2D1) x TM-1
F01	F1	Inter - GH x A2D1	2(A2D1) x TM-1
G01	G.lon	*G. longicalyx*	Stelly Lab
H01	G.lon	*G. longicalyx*	Stelly Lab
A02	G.arm	*G. armourianum*	USDA D2-1-6
B02	G.arm	*G. armourianum*	USDA D2-1-6
C02	G.arb	*G. arboreum*	USDA A2-49-1
D02	G.arb	*G. arboreum*	USDA A2-49-1
E02	HLA	Tri-species Hybrid	FADD - GH, GL, Garm
F02	G.rai	*G. raimondii*	Stelly Lab
G02	Water	-	Non Template Control
H02	Water	-	Non Template Control

*GH, GL, Garb, and Garm represent *G. hirsutum*, *G. longicalyx*, *G. arboreum*, and *G. armourianum* respectively.

Figure 2.3 KASP marker screening panels. Screening panels containing control and mapping samples used for determining successful and unsuccessful markers via KASP assay genotyping. **(A.)** Panel used for screening markers derived from *G. barbadense*, "*G. barbadense* screening panel". **(B.)** Panel used for screening markers derived from *G. tomentosum* and *G. mustelinum*. **(C.)** Panel used for screening markers derived from *G. longicalyx*. **(D.)** Panel used for screening markers derived from *G. armourianum*.

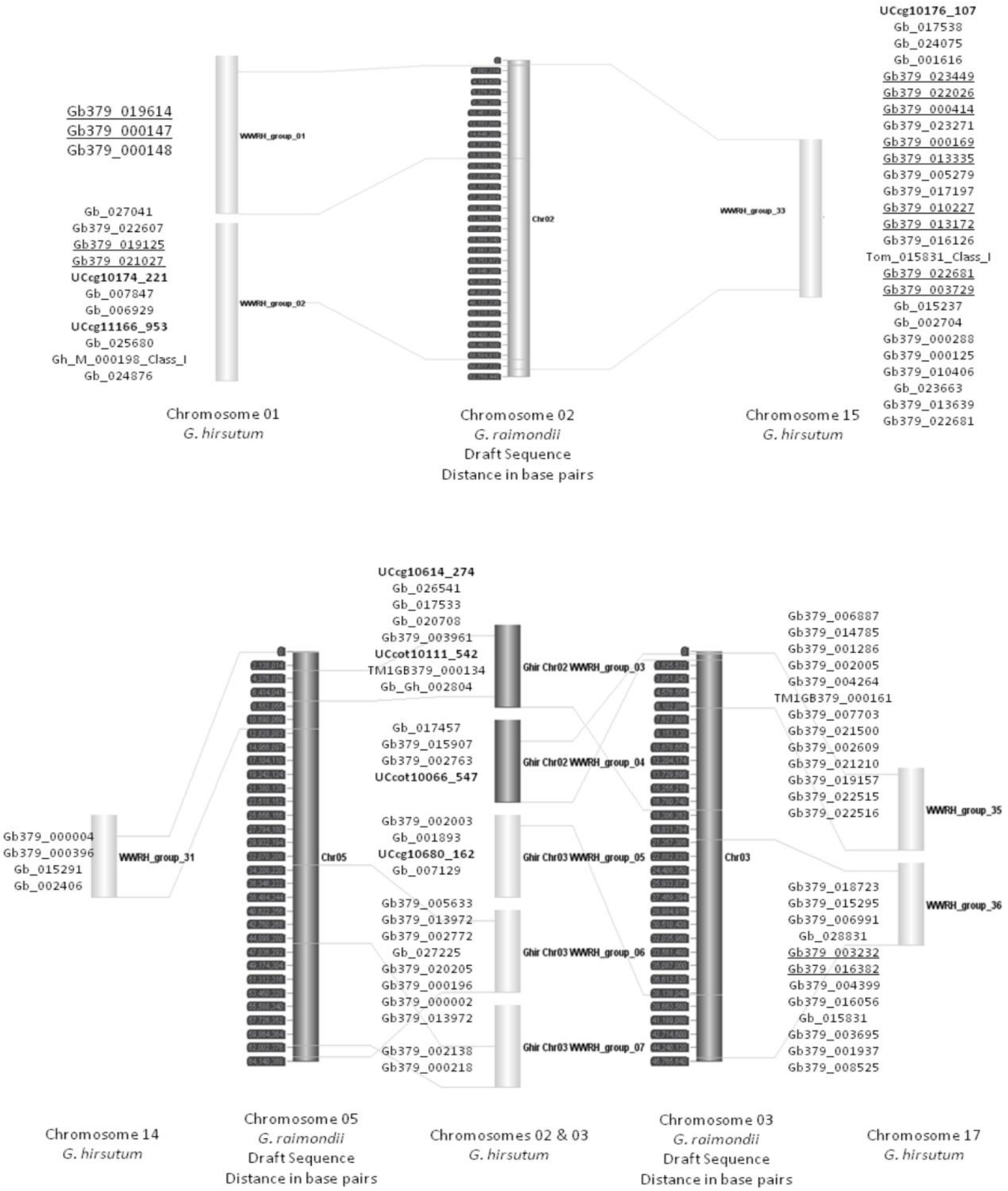


Figure 2.4 Wide-cross whole-genome radiation hybrid bin map. Wide-cross whole-genome radiation hybrid map generated from genotypes of 131 irradiated F1 (*G. hirsutum* line TM-1 × *G. barbadense* cv. 3–79) individuals in Carthagene using LOD score of 3. Bins consist of all markers which fall in a single syntenic group as determined by Carthagene. Bins are aligned to the *G. raimondii* (D_5) draft genome sequence by BWA mapping of individual SNP markers. Bold markers indicate markers from the Van Deynze et al. [20] data set that were mapped in the Yu et al. [14] paper. Underlined markers indicate markers for which the sequences were overlapped.

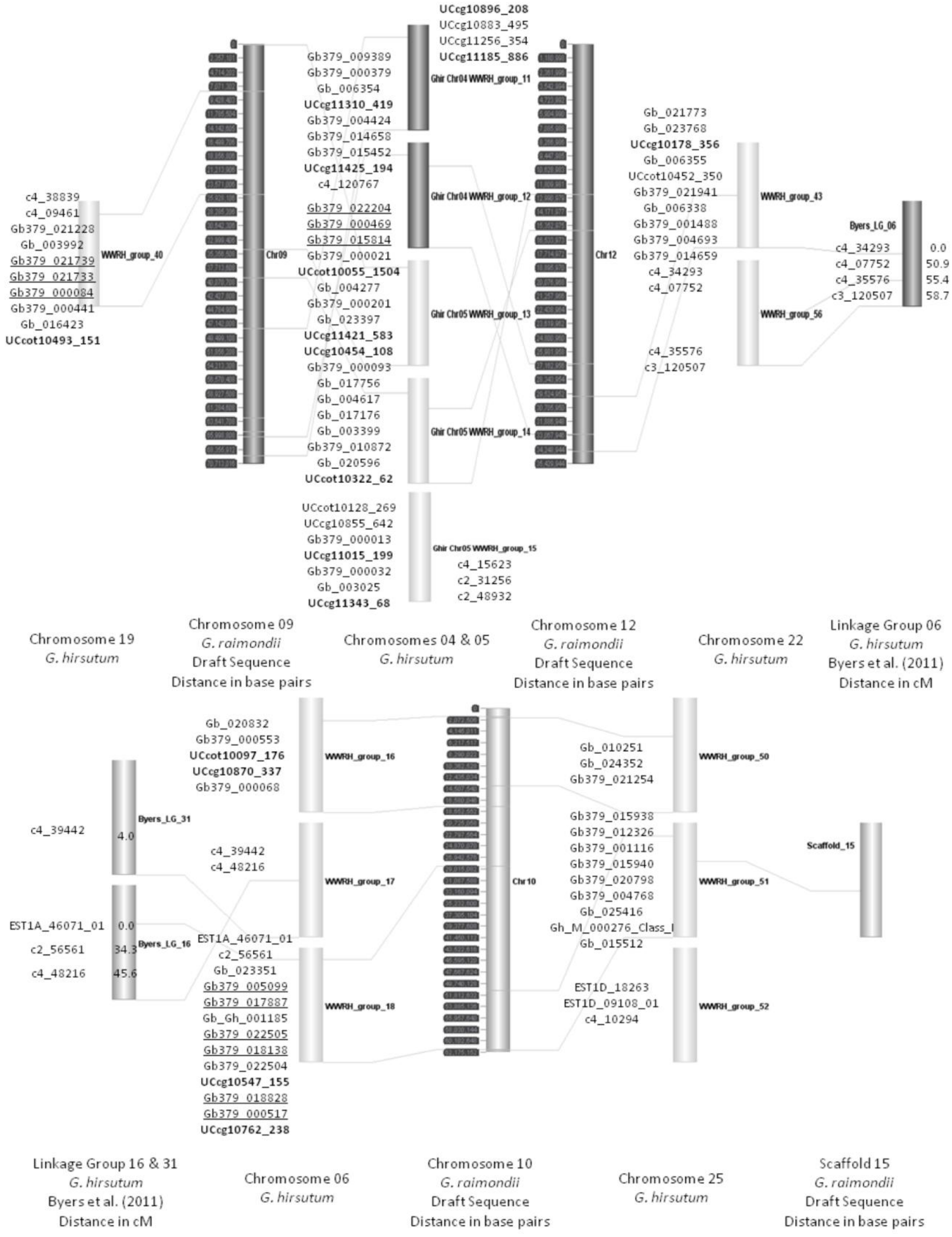


Figure 2.4 (continued)

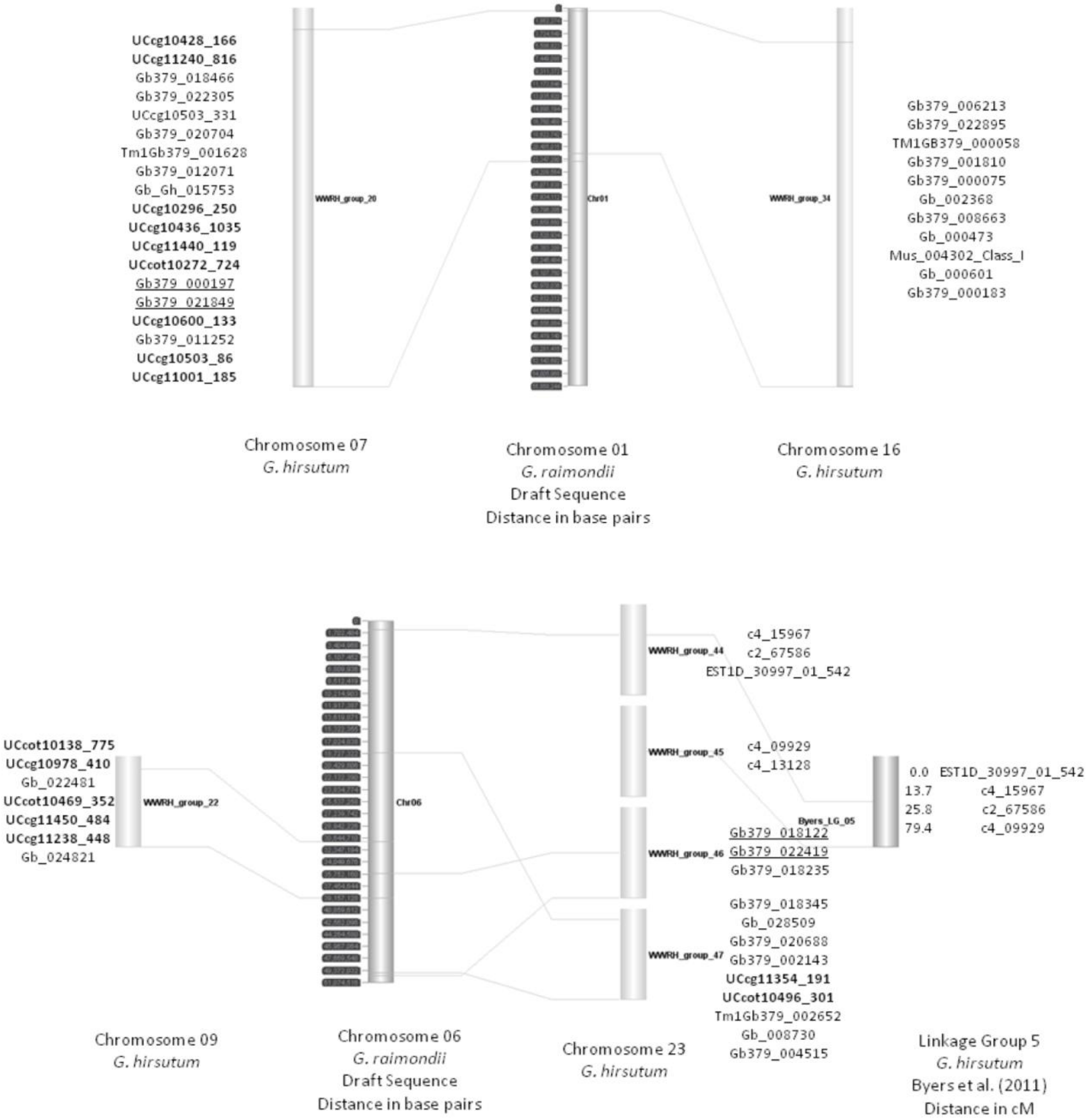


Figure 2.4 (continued)

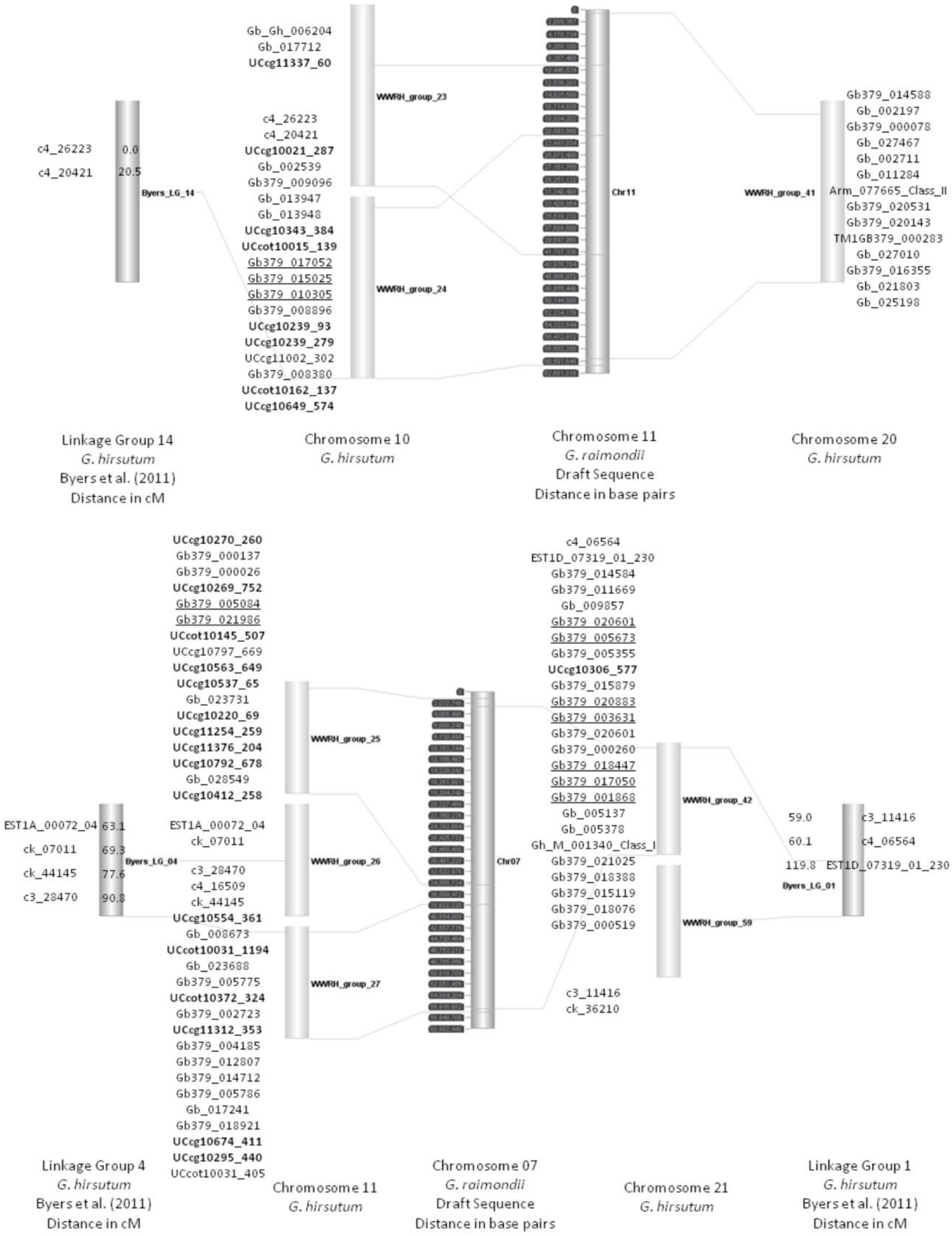


Figure 2.4 (continued)

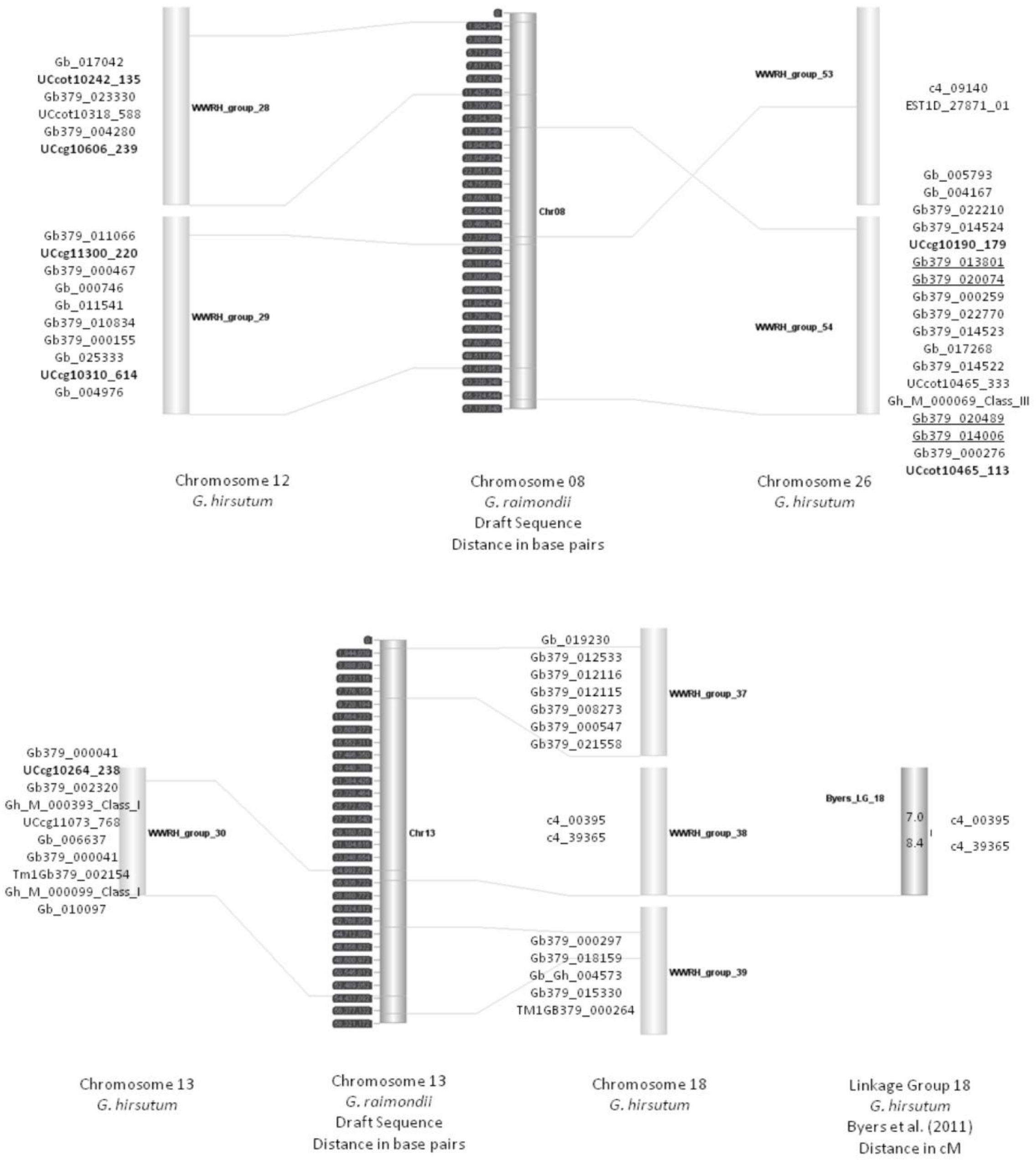


Figure 2.4 (continued)

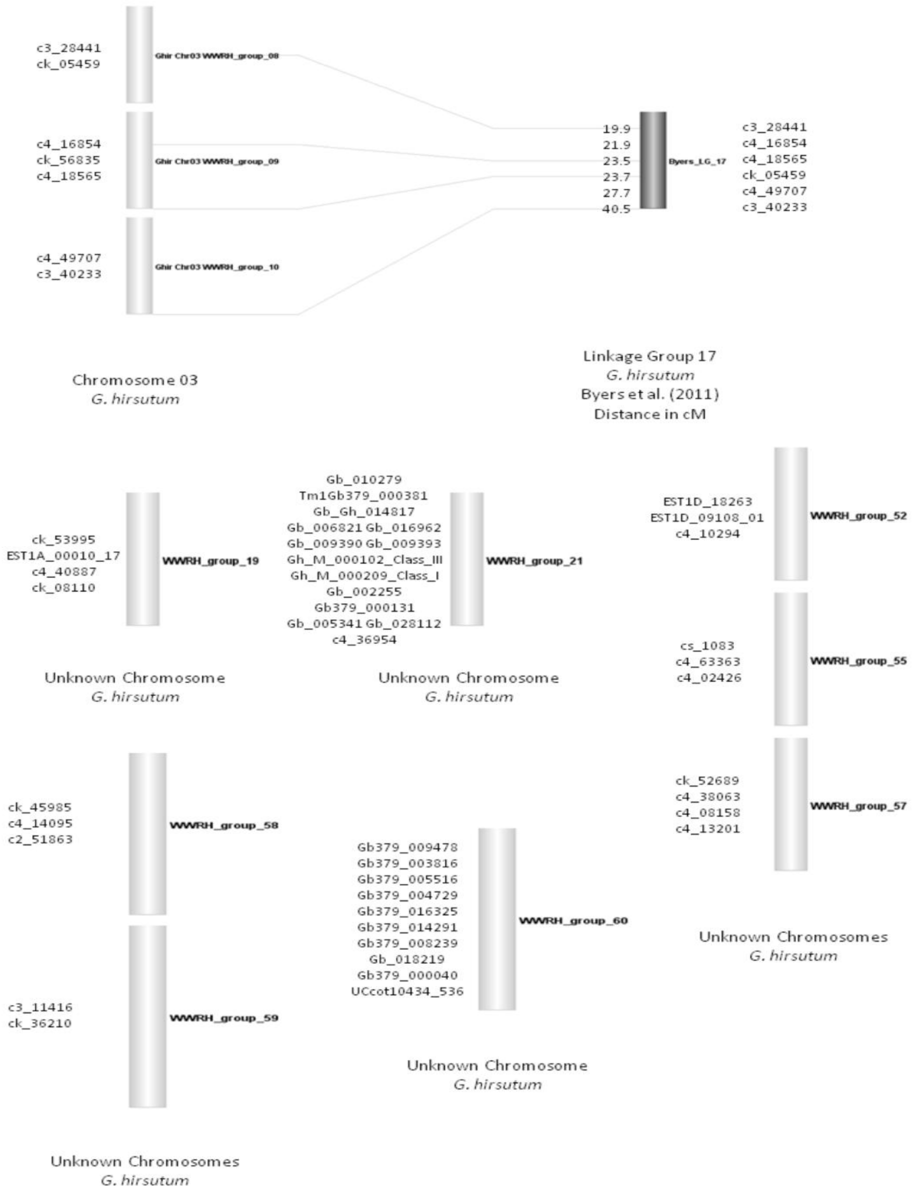


Figure 2.4 (continued)

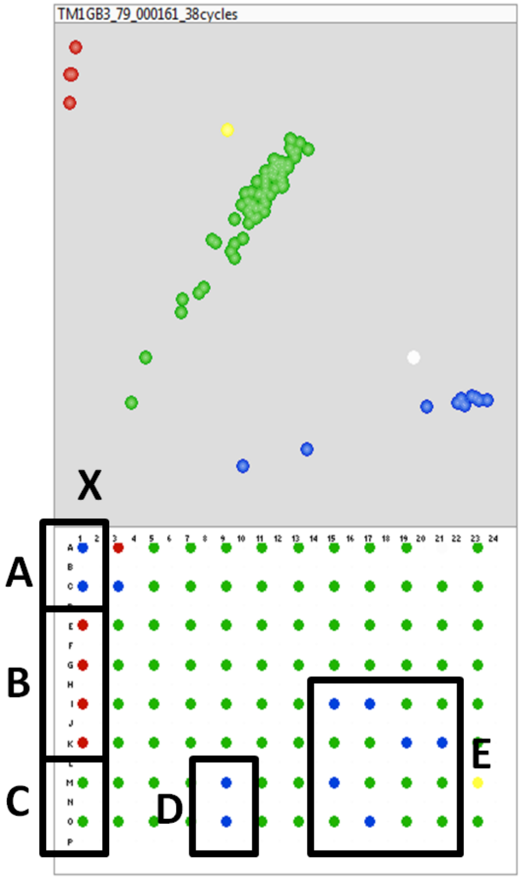


Figure 2.5 KASP genotyping of marker UCcg10563_649. A1 quadrant of a 384-well plate KlusterCaller image of genotyping KASP assay for marker UCcg10563_649 after 38 cycles. **A.)** *G. barbadense L.* $(AD)_2$ cv. 3–79, **B.)** *G. hirsutum L.* $(AD)_1$ line TM-1, **C.)** F_1 euploid hybrid *G. barbadense L. (*$(AD)_2$ cv. 3–79 and *G. hirsutum L.* $(AD)_1$ line TM-1, **D.)** F_1 hypo-aneuploid lines for AD-Chromosome 17, **E.)** Wide-cross whole-genome radiation hybrid samples, **X.)** Non-template (negative control), in B2 quadrant (not shown). Both samples in D and 6 of the samples in E show deletions due to the shift in genotype from F_1-green to homozygous for the *G. barbadense L. (*$(AD)_2$ allele – blue. [The two samples shown in white and yellow were considered questionable genotypes as they did not fall directly in a cluster].

Deletion analysis of G. barbadense SNPs

The number of WWRH deletions was significantly higher for only 6 (1.2%) of the 509 *G. barbadense* markers used for WWRH mapping, based on Tukey's boxplot method of outlier detection. These markers were Gb379_011066, Gb379_000467, UCcg10762_238, UCcg10614_274, c4_101926, and c4_44618.

Functional analysis

A total of 117 contigs containing 118 SNPs (three of the shared SNPs in **Figure 2.1** are loci which have three alleles so these were not included in the analysis) were found to be shared between all wild species relative to the TM-1 *G. hirsutum* reference. When translation was predicted using AUGUSTUS software (http://bioinf.uni-greifswald.de/augustus/) (Keller, Kollmar et al. 2011) and *Theobroma cacao* as the model species, 52 out of the 116 translations were found to generate different amino acid sequences or non-synonymous substitutions, which corresponds to a non-synonymous to synonymous ratio (N/S) of 0.8125. As a method of comparison to the first set, a random set of 118 SNPs from the overall non-redundant Class I and II data set was chosen. These SNPs represented 118 different contigs. When the same analysis using AUGUSTUS was performed with this random set of contigs, 33 out of the 109 translations were found to generate different amino acid sequences, N/S of 0.4342.

Discussion

Interspecific germplasm introgression provides a powerful way of introducing novel beneficial alleles into breeding germplasm of Upland cotton. Its use has been constrained by the long time periods required for introgression and the difficulty of breaking linkage blocks. But when patience is exhibited and recombination has occurred to break linkage blocks to allow for introgression of interspecific segments, highly beneficial products can be obtained. Such is the case with BARBREN which was created to move reniform resistance found in *G. barbadense* into *G. hirsutum* along with superior fiber characteristics (Bell, Quintana et al. 2013). Combining features of *G. barbadense* with *G. hirsutum* has been a long-standing desire, because *G. barbadense* offers many superior fiber trait characteristics but does not produce the high lint yield of traditional *G. hirsutum* cultivars. The large number of *G. barbadense* SNPs identified in this study, 10,521, and particularly the 8,307 Class I and Class II markers will provide markers for a large number of genes in which differences exist between *G. hirsutum* and *G. barbadense*. A modest number, 509 markers, have been validated (from 594 tested) and the majority of these have been anchored to an allotetraploid chromosome by WWRH mapping, therefore relative physical location is known (**Figure 2.4**). A larger number of markers and/or larger WWRH panel would have been needed to link all of the singleton markers into the syntenic groups. Markers within each group were close enough for statistical association, but the number and variety of deletions was insufficient to accurately determine physical order. The average number of plants with a deletion for a given SNP was only 8.9 (6.9%), far short of the optimal deletion rate i.e., 50%. Accurate ordering would require a far larger population of similarly irradiated plants or a population with a much higher deletion rate.

Additional investigation into patterns of deletions showed that only 6 (1.2%) of markers across individuals had a statistically significant number of deletions. This was a statistically insignificant number from the overall set. The relative abundance of these marker deletions suggests that the respective chromosomal segments have higher propensities for deletion and/or post-occurrence recovery of induced deletions. Some chromosomal segments may be more likely to be lost after pollen irradiation, and/or there may be significant differences in selection for/against loss of specific genes/alleles in these segments that correspond to the identified markers.

Placement of the syntenic groups relative to the D_5 reference sequence revealed most to be in non-pericentromeric and non-telomeric regions, i.e., similar to the pattern observed for individual markers (**Figure 2.2**). SNPs were distributed unevenly across the chromosomes. A bimodal distribution was observed for each pseudo-chromosome scaffold, with large numbers of SNPs near the subtelomeric regions and small numbers in centromeric and telomeric regions, as would be expected given the metacentric nature of cotton chromosomes and the fact that the markers were derived from expressed sequences (Luo, Mach et al. 2012). In addition all SNPs were found to integrate into a single syntenic group, unlike SSRs which integrate across groups, which implies the majority of SNPs are subgenome specific and will be useful for breeding as identifying a unique position in the genome.

Mapping experiments were focused on *G. barbadense*. Like Upland cotton, it is a cultivated species and represents approximately five percent of the worldwide cotton production. However, *G. barbadense* being not as highly improved as *G. hirsutum*, it retains some alleles that are deleterious when brought into a *G. hirsutum* background. Some research has been done utilizing chromosome substitution lines, in which a single chromosome in the *G. hirsutum* allotetraploid has been replaced with the same allotetraploid chromosome from a different species, for example *G. barbadense* (Stelly, Saha et al. 2005). Many beneficial regions have been identified for yield components as well as fiber quality traits (Jenkins, Wu et al. 2006; Jenkins, McCarty et al. 2007). Recombinant inbred lines from these chromosome substitution lines have also been generated recently that will assist in introgression efforts once markers from a high-density dataset such as was developed here have been located for target areas. It has been shown that cryptic beneficial alleles are typically masked in the overall *G. barbadense* background (Saha, Wu et al. 2013).

Like *G. barbadense*, the other wild allotetraploid species, *G. mustelinum* and *G. tomentosum*, included in this study have also been shown to host cryptic beneficial alleles. These species have been integrated into the chromosome substitution line development effort and will provide additional trait resources for movement into a *G. hirsutum* background. Being of allotetraploid genome constitution, most genomic segments from these species will easily be moved into *G. hirsutum.* Integrating genes from wild diploid species like *G. longicalyx* and *G. armourianum* is much more complicated, because their diploid genomes are vastly different from the *G. hirsutum* genome. However with the longer divergence time and large number of diploid species available, (~45) compared to uncultivated tetraploid species (~3-5), a much larger number of unique beneficial alleles may be found in diploid *Gossypium* species. Inventive methods of creating synthetic polyploids with diploid species have been devised to facilitate transfer of genetic material into *G. hirsutum*, as was the case for *G. longicalyx*, in order to create Upland cottons with strong resistance to reniform nematodes (Bell, Forest Robinson et al. 2014). Two sister lines with strong reniform nematode resistance, LONREN-1 and LONREN-2, were released but subsequently discovered to suffer early growth season "stunting" suggesting a possible linkage drag or pleiotropic effect (Zheng 2012). In general, practical utilization of introgressed alien germplasm demands precise genetic manipulation to separate linked beneficial and deleterious alien genes, for which numerous markers are essential. Thus, large numbers of markers are needed for each germplasm source, such as the markers developed here. Class I markers will be exceptionally useful as they can be used for determining haplotype information being the only marker identified within a contig.

Studying shared markers between the different species of varied genome composition relative to cultivated *G. hirsutum* can be used to deduce the theoretical ancestral allele at a locus, as well as to suggest functional properties of a locus. The distribution of shared markers is depicted in **Figure 2.1**. For markers at which all wild species share a common allele but *G. hirsutum* differs, it can be inferred that the wild species share the ancestral allele and *G. hirsutum* contains an alternative allele. Such alleles are good candidates for being functionally important to domestication, cultivation or agronomic performance, or in linkage disequilibrium with such genes. The non-synonymous to synonymous ratio was found to be much higher in the 118 SNPs shared between species (0.8125) than in 118 randomly selection SNPs from the final set (0.4342). This implies that there is a stronger positive selection upon the SNPs where *G. hirsutum* has an allele which is non-ancestral. This further supports the hypothesis that these loci are likely to be important in beneficial traits in *G. hirsutum*.

Some markers within the same species were identified from multiple scaffolds, but were identical in SNP and flanking sequence (**Additional files 2.3, 2.5, 2.7, 2.9, 2.11**). This is likely due to genes from gene families which exhibit very high levels of sequence similarity. Therefore multiple hits are expected for genes that have expanded in the TM-1 or *G. hirsutum* lineage or are duplicated within a genome (paralogs). Another possible explanation is that these hits relate to contigs that contain different isoforms of the same gene. When SNPs derived from multiple isoforms are mapped physically or by linkage, they will locate to a single locus. SNPs from the former case will occupy multiple locations in which different members of the gene family are found. We classified these SNPs separately as they may represent multiple loci throughout the genome and present another level of difficulty for genotyping. The diploid species were found to have many more SNPs exhibiting within-species redundancy, which are marker sequences that are identical but generated from multiple contigs in the assembly, than the tetraploid species. This result was expected because the tetraploid *G. hirsutum* that was utilized as a reference likely received a copy of each gene from ancestors of the A and D subgenomes during polyploidization. Thus relative to each diploid, the reference would have twice as many copies for each gene (assuming no gene loss or duplication has occurred, or co-assembly) and would lead to derivation of the same SNP sequence from each of the homeologous copies found in *G. hirsutum* when analyzing the diploid species.

CHAPTER III

BAC-END SEQUENCE-BASED SNP MINING IN ALLOTETRAPLOID COTTON (*GOSSYPIUM*) UTILIZING RESEQUENCING DATA, PHYLOGENETIC INFERENCES AND PERSPECTIVES FOR GENETIC MAPPING

Introduction

Marker development in crop species has been an important aspect to facilitate genomics-based crop improvement. Single nucleotide polymorphisms (SNPs) are the most abundant form of markers in all organisms, as they have a possibility of occurring at every nucleotide position in the genome. This makes them ideal candidates for construction of high-density genetic maps, which can then be used for marker-based crop improvement and genetic analyses. In general, for marker-assisted selection, a large number of molecular markers are required in crop species as the majority of the traits of interest such as yield, drought and heat tolerance, nitrogen and water use efficiency, disease resistance, and fiber quality are complex and controlled by many genomic loci of small effect. Therefore, for marker-assisted breeding to be effective, markers for most of the regions controlling a trait need to be included in selection criteria (Mammadov, Aggarwal et al. 2012). SNPs have been found to occur approximately every 60-120 bp in the maize genome (Ching, Caldwell et al. 2002), every 268 bp in the rice genome (Shen, Jiang et al. 2004) and every 185-266 bp in the soybean genome (Lam, Xu et al. 2010). The markers developed to-date in cotton have been limited mostly to amplified fragment length polymorphisms (AFLPs), restriction fragment length polymorphisms (RFLPs) and simple sequence repeats (SSRs). Some recent efforts to identify SNPs in cotton using transcriptome sequencing in intraspecific and interspecific cotton lines (Hulse-Kemp, Ashrafi et al. 2014; Zhu, Spriggs et al. 2014), single copy sequences between *G. hirsutum* and *G. barbadense* (Van Deynze, Stoffel et al. 2009) and a variety of reduced representation libraries (RRLs) techniques investigating combinations of *G. hirsutum* lines (Byers, Harker et al. 2012; Rai, Singh et al. 2013; Gore, Fang et al. 2014; Islam, Thyssen et al. 2014; Zhu, Spriggs et al. 2014) have been reported and have been increasing in efficiency using a diverse range of germplasm. These studies in cotton have generated anywhere from a few hundred SNPs to tens of thousands of SNPs. However success in cotton has been limited compared to crop species such as maize (Elshire, Glaubitz et al. 2011; Hansey, Vaillancourt et al. 2012), soybean (Hyten, Cannon et al. 2010), barley and wheat (Poland, Brown et al. 2012), which also used a broad range of germplasm and have been able to develop hundreds of thousands of SNPs.

Reasons for reduced success with SNP development within *Gossypium* compared to efforts in other crops is perhaps due to the evolutionary history of the cotton genome and recent polyploid generation. The most widely cultivated species of cotton, *Gossypium hirsutum* L., descends from a recently formed allotetraploid hybrid (1-2 MYA), $2n = 4x = 52$, genomic formula $2[AD]_1$ (Wendel, Brubaker et al. 2009). Modern cultivated cotton, including Upland types, have undergone at least two independent bottleneck events, domestication, and elite cultivar selection, which further reduced the overall diversity that can be found in elite cultivars. The ancestral A- and D-like genomes are thought to have diverged only 5-10 MYA and polyploidization of G. hirsutum 1-2 MYA. Due to limited time for evolutionary divergence, homeologous regions of the A- and D-subgenomes of cotton retain a very high similarity. While reference sequences have recently become available for the diploid cottons, *G. raimondii* (extant relative to the allotetraploid D-subgenome) (Paterson, Wendel et al. 2012) and *G. arboreum* (extant relative to the allotetraploid A-subgenome) (Li, Fan et al. 2014), a reference is not yet available for allotetraploid *G. hirsutum*.

Large-scale SNP development in cotton has been hindered by the lack of a high-quality *G. hirsutum* reference which, in concert with high levels of similarity between subgenomes and low levels of diversity, make it very difficult to unambiguously map short-read sequences obtained from next-generation sequencing technologies (Zhu, Spriggs et al. 2014). Recently we developed three independent bacteria artificial chromosome (BAC) libraries, two generated from restriction-enzyme partial digestion (BstYI/HindIII) and one generated by random-shearing (Saski et al. unpublished). From these, 179,209 high quality BAC-end sequences (BESs) were generated. Development of markers aligned to BAC-end sequences is desired for several reasons: 1.) The quality of Sanger-sequencing is currently the highest on a single read basis, 2.) The markers can serve as a rigid interface between a BAC-based physical map with genetic maps, and 3.) Taken together, long-range contiguity and marker distribution can begin to be contextualized on a genome-wide scale. Recent reports suggest the utility of BESs in simple sequence repeat (SSR) marker development in pigeon pea (Bohra, Dubey et al. 2011), cotton (Frelichowski, Palmer et al. 2006), and peanut (Wang, Penmetsa et al.

2012) and for SNP development via PCR-directed sequencing methods in apple (Han, Chagne et al. 2009) and in *Citrus* (Ollitrault, Terol et al. 2012) to anchor genomic sequences.

Thus, the primary goals of the present study were: (i) to develop large numbers of genomic-based SNPs for cultivated cotton, *G. hirsutum*; (ii) to compare the levels of diversity among elite cultivars, a wild accession, an additional tetraploid species (*G. barbadense* L.) and a diploid cotton species (*G. longicalyx*); (iii) experimentally validate *in silico*-derived SNPs; and (iv) demonstrate mapping ability and utility of developed SNPs. This paper reports on utilizing BESs as a high-quality reference for SNP discovery through resequencing. The new SNPs will be a resource for SNP-based integration of the physical and genetic maps and the methodology can also serve as a useful model in resolving other complex plant genomes.

Materials and Methods

Source of BAC clones and BAC-end sequencing

Two complementary BAC libraries (*Bst*YI and *Hind*III) and a random sheared BAC library from *G. hirsutum* genetic standard line Texas Marker-1 (TM-1) were used in this study. BAC DNA was isolated and sequenced by Sanger methods at USDA-ARS (Stoneville, Mississippi). A total of 179,209 BAC-end sequences were retained after quality trimming and filtering (LIBGSS_039228) (Saski et al. unpublished).

Plant material and DNA extraction

The seed of 11 *G. hirsutum* lines (TM-1, Sealand 542, PD-1, Paymaster HS-26, M-240 RNR, Fibermax 832, Coker 312, SureGrow 747, Stoneville 474, Tamcot Sphinx, and TX0231) and the *G. barbadense* genetic standard line, 3-79, were planted at Texas A&M University, UC Davis or the USDA-ARS facility in New Orleans, USA. Young leaf tissues were sampled from each plant and used to isolate genomic DNA using the Qiagen DNeasy (Qiagen, Valencia, USA) plant extraction kit following manufacturer instructions, including RNase digestion. DNA concentrations and qualitative absorbance values were determined using the Nanodrop Spectrophotometer (Thermo Fisher Scientific, Waltham, USA).

Library Preparation and Sequencing

Double-stranded DNA was quantified with a PicoGreen assay on the Synergy HT plate reader (Bio-Tek, Highland Park, USA)). Libraries were prepared using an in-house protocol with individual barcoding. One and a half micrograms of each DNA sample was randomly sheared using a Bioruptor instrument (Diagenode, Denville, USA) and then size-selection was performed using AMPure XP beads to 300-500 bp. Fragments were end-repaired using NEBNext End Repair Module and then purified with AMPure beads. Next an adenine was ligated at the end of the fragments, adapters were ligated and the final products were purified with AMPure beads. Enrichment- PCR was performed for 14 cycles then the PCR product was run on a 1.5% gel to confirm enrichment of product and size range. A final round of purification was performed using AMPure beads. The final libraries were assessed for quality with the Bioanalyzer (Agilent, Santa Clara, USA) to determine final library size and concentration. Each sample was sequenced with paired-end sequencing (2x100 bp) on two Illumina HiSeq2500 lanes. Raw, paired-read sequence files were uploaded to NCBI under BioProject PRJNA257223 and SRA numbers (SRX667500 - *G. hirsutum* TM-1, SRX668168 - *G. hirsutum* Sealand 542, SRX668322 - *G. hirsutum* PD-1, SRX668354 - *G. hirsutum* Paymaster HS-26, SRX669467 - *G. hirsutum* M-240 RNR, SRX669468 - *G. hirsutum* Fibermax 832, SRX669469 - *G. hirsutum* Coker 312, SRX669470 - *G. hirsutum* SureGrow 747, SRX669471 - *G. hirsutum* Stoneville 474, SRX669472 - *G. hirsutum* Tamcot Sphinx, SRX669473 - *G. hirsutum* TX0231, SRX669474 - *G. barbadense* 3-79). Raw reads for *G. hirsutum* Acala Maxxa and for *G. longicalyx* were obtained from the NCBI Small Read Archive under numbers SRR617482 and SRR617704.

SNP mining from G. hirsutum aligned to BAC-end sequences

All raw sequence files were assessed for initial quality using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and the first 13 bases from each read were removed due to poor quality. The remaining reads were quality trimmed, adapters and any reads fewer than 40 bases were removed using the FastX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). The files were assessed for final quality and concatenated into a single file to be utilized as single-end read data, due to the size of the BESs being used as a reference. Once the sequences had been processed they were imported into CLC Genomics Workbench v 6.0.2 (Valencia, USA). Reads from Sealand

542, PD-1 and 3-79 were aligned to the BAC-end sequence reference using different length and similarity fractions in order to determine optimal parameters. Iteration-1 utilized length fraction 0.70 and similarity fraction 0.99, while Iteration-2 utilized length fraction 0.99 and similarity fraction 0.98. SNPs were called using the Probabilistic variant caller in CLC Genomics Workbench using minimum sequence depth of six, variant probability of 50.0, required presence in both forward and reverse reads, four maximum expected variants and one standard genetic code. Theoretical homeo-SNP positions (variants between homeologous regions caused by ambiguous mapping of homeologous reads from different subgenomes in the allotetraploid) were determined by aligning the TM-1 Illumina-based sample, which is the same genotype as the reference, to call SNPs using the same parameters as all other samples. This identified homeo-SNP positions were removed from the SNPs identified in the other samples.

A random set of called 32 SNP positions was manually viewed to assess quality of alignments around the SNP positions, which included (i) markers only found in Iteration-1; (ii) markers only found in Iteration-2; and (iii) markers found in both Iteration-1 and Iteration-2 from the SNPs discovered using *G. barbadense* 3-79 were selected for empirical testing. Half of the 32 markers were identified as overlapping transcriptome-derived SNPs as reported in Hulse-Kemp et al. (2014), whereas half of the SNPs did not overlap the Hulse-Kemp et al. dataset and thus were assumed to not be associated with the transcriptome. Primers were designed for KASP (LGC Genomics) SNP assays using BatchPrimer3 (allele-specific primers & allele-flanking primers, Tm - optimum: 57ºC, minimum: 55ºC, maximum: 60ºC, max difference: 2ºC product size-minimum: 50bp, optimum: 50bp, maximum: 100bp). Primers were synthesized and diluted according to KASP developer (LGC Genomics, Hoddesdon, UK) instructions. KASP assays were run on a "*G. barbadense* screening panel" containing 12 samples including TM-1 (Stelly Lab), TM-1 (USDA), 3-79 (x2), F1 – 3-79xTM-1 (x2), RIL01-04 (3-79xTM-1) and water non-template control (x2) according to manufacturer's suggested PCR conditions. Plates were read using the Pherastar at 38, 44 and 50 cycles, and then analyzed using the KlusterCaller program. SNP assays were labelled as "good" if samples produced clean clusters that allowed for differentiation and scoring of the two parents and the F1 genotypes, or "bad" if no definable clusters were produced or no amplification occurred. (These definitions of "good" and "bad" will be used subsequently throughout the rest of the manuscript.) Markers included in the "good" and "bad" categories for Iteration-1 and 2 were calculated (**Table 3.1**) to determine optimal parameter settings.

Table 3.1 List of unfiltered SNPs determined for all species.

Type	Marker Name	KASP Result	Identified in *G. barbadense* Mapping		Homeo-SNP Removal		Final Result	
			Iteration 1	Iteration 2	Iteration 1	Iteration 2	Iteration 1	Iteration 2
Not Transcriptome Associated	GH_TBb001A07f_381	GOOD	YES	YES	RETAIN	RETAIN	GOOD	GOOD
	GH_TBb001A07f_486	BAD	NO	YES	-	REMOVE	-	-
	GH_TBb001A23r_531	BAD	NO	YES	-	RETAIN	-	BAD
	GH_TBb001A23r_614	GOOD	YES	YES	RETAIN	RETAIN	GOOD	GOOD
	GH_TBb001D22r_204	GOOD	YES	YES	RETAIN	RETAIN	GOOD	GOOD
	GH_TBb001D22r_445	BAD	YES	NO	RETAIN	-	BAD	-
	GH_TBb001D22r_511	BAD	YES	YES	RETAIN	RETAIN	BAD	BAD
	GH_TBb001B05f_180	GOOD	YES	YES	REMOVE	REMOVE	-	-
	GH_TBb001B05f_564	BAD	YES	YES	REMOVE	REMOVE	-	-
	GH_TBb001C03f_116	GOOD	NO	YES	-	RETAIN	-	GOOD
	GH_TBb001C03f_401	GOOD	YES	YES	RETAIN	RETAIN	GOOD	GOOD
	GH_TBb001F01r_117	BAD	YES	YES	REMOVE	REMOVE	-	-
	GH_TBb001F01r_310	BAD	YES	NO	REMOVE	-	-	-
	GH_TBb001A17r_218	BAD	YES	YES	REMOVE	REMOVE	-	-
	GH_TBb001F06f_303	BAD	NO	YES	-	REMOVE	-	-

Table 3.1 (continued)

Type	Marker Name	KASP Result	Identified in *G. barbadense* Mapping		Homeo-SNP Removal		Final Result	
			Iteration 1	Iteration 2	Iteration 1	Iteration 2	Iteration 1	Iteration 2
Transcriptome Associated	GH_TBb004J20r_76	GOOD	YES	YES	REMOVE	REMOVE	-	-
	GH_TBb004J20r_348	GOOD	YES	YES	RETAIN	RETAIN	GOOD	GOOD
	GH_TBb053N14f_270	GOOD	YES	YES	RETAIN	RETAIN	GOOD	GOOD
	Gh_TBh036B20r_583	GOOD	NO	YES	-	RETAIN	-	GOOD
	GH_TBr162H20f_547	BAD	YES	YES	REMOVE	REMOVE	-	-
	GH_TBb046O02r_64	BAD	YES	YES	REMOVE	REMOVE	-	-
	GH_TBb046O02r_138	BAD	YES	YES	REMOVE	REMOVE	-	-
	GH_TBb046O02r_418	BAD	YES	YES	REMOVE	REMOVE	-	-
	GH_TBb069A06f_248	BAD	YES	NO	RETAIN	-	BAD	-
	GH_TBb069A06f_304	BAD	YES	YES	REMOVE	REMOVE	-	-
	GH_TBh034D07r_276	GOOD	YES	YES	RETAIN	RETAIN	GOOD	GOOD
	GH_TBh030P12r_332	BAD	YES	YES	REMOVE	REMOVE	-	-
	GH_TBb119O19f_465	MAYBE	YES	YES	RETAIN	RETAIN	N/A	N/A
	GH_TBh055K17f_57	BAD	YES	YES	REMOVE	REMOVE	-	-
	GH_TBh055K17f_506	BAD	YES	YES	REMOVE	REMOVE	-	-
	GH_TBh023O21f_162	GOOD	YES	NO	REMOVE	RETAIN	-	GOOD
	GH_TBh023O21f_537	GOOD	YES	YES	RETAIN	RETAIN	GOOD	GOOD

All 12 *G. hirsutum* samples and the *G. barbadense* sample were analyzed using Iteration-2 parameters. The *G. longicalyx* sample was analyzed using the same length fraction but similarity fraction of 0.96. Samples were processed individually and homeo-SNPs identified using TM-1 were removed. Unique SNP positions were collated into a master file and required 100% homozygous identification in at least one *G. hirsutum* line. Positions where coverage was one to two standard deviations above the average coverage in the homozygous sample(s) were removed, as regions with high coverage are likely to be associated with repetitive regions. Additional unique SNP positions identified as being homozygous in *G. barbadense* or *G. longicalyx* were added to the master file. Only positions with probability scores >0.98 were retained for *G. barbadense* and *G. longicalyx.* A hierarchical method was utilized to determine retention; SNPs were retained in *G. hirsutum* file initially, then *G. barbadense* and then *G. longicalyx*. Three non-redundant variant call format (VCF) files were obtained from each of the three species groups. These VCF files were uploaded to NCBI's dbSNP.

A final VCF file was derived using GATK software (https://www.broadinstitute.org/gatk/) that contained genotypes at all positions for all samples using the list of non-redundant SNP positions, binary alignment map (BAM) files exported from CLC Workbench for each individual sample, and BES FASTA file. The GATK UnifiedGenotyper command with default settings except for the GENOTYPE_GIVEN_ALLELES, EMIT_ALL_SITES and minimum base quality of 20 options were used to call genotypes for all previously identified SNP positions along the BES reference for the 12 *G. hirsutum* lines, *G. barbadense*, and *G. longicalyx*.

Distributions of SNP types, number of missing calls per sample and number of heterozygous calls per sample were calculated from the final VCF file using VCFtools (http://vcftools.sourceforge.net/). The number of homozygous differences between samples was determined using BCFtools command gtcheck.

Marker alignment to diploid cotton reference genomes

Markers identified were exported with 50-bp flanking sequence on both sides. SNPs were coded with IUPAC ambiguity codes and the total 101 bp sequence of each SNP with flanking sequence was mapped using Burrows-Wheeler Aligner (BWA) in Galaxy (Goecks, Nekrutenko et al. 2010) using default settings. All SNPs were mapped to both the *Gossypium raimondii* (D_5) reference genome (Paterson, Wendel et al. 2012) and the *Gossypium arboreum* (A_2) reference genome (Li, Fan et al. 2014). Alignments were corrected to report SNP positions.

SNP screening

A set of 48 randomly selected intra-specific SNP markers were selected for experimental screening to estimate average validation rate of *in silico* determined intraspecific SNPs. Primers were designed using previously mentioned parameters in BatchPrimer3, and KASP assays were run per manufacturer instructions. A "*G. hirsutum* screening panel" that included 32 *G. hirsutum* lines, *G. barbadense* line 3-79 and water non-template controls was used (**Additional File 3.1**). An additional set of 96 markers developed from the 3-79 line were selected based on diploid (D_5 genome) alignment information with inferred positioning on allotetraploid chromosomes 12 and 26 (Blenda, Fang et al. 2012). The markers were designed and screened on the *G. barbadense* screening panel.

A random set of 32 markers developed from *G. longicalyx* were selected and primers were designed. KASP assays were run per manufacturer's instructions on a "*G. longicalyx* screening panel" which contained *G. longicalyx* (x2), *G. hirsutum* cv. TM-1 (x 2), synthetic allotetraploid "FADD" (x 2), 2 BC1F1 samples (FADD x TM-1), *G. barbadense* line 3-79, *G. hirsutum* accession TX0231 and 2 non-template controls. Plates were cycled and analyzed as previously mentioned. SNP sequences and primers for all screened markers are listed in **Additional File 3.2**.

Interspecific linkage mapping

Good markers obtained in screening of the *G. barbadense* SNPs were used to genotype 118 F2 (TM-1x3-79) individuals, two parents (*G. hirsutum* line TM-1 and *G. barbadense* line 3-79) and F1 (3-79xTM-1) as controls. KASP assays were run using the Fluidigm system in 96.96 dynamic array format, utilizing multiple arrays, and read using the Fluidigm BioMark HD (Fluidigm, San Francisco, USA). Clustering for genotyping was performed using Fluidigm SNP Genotyping Analysis software. Genotypes of the 118 F2s for successfully genotyped markers were imported into JoinMap 4.1 (JW 2011), identical markers were removed and linkage mapped using the maximum likelihood algorithm and Haldane's mapping function with default parameters. Linkage groups were determined using LOD scores of 5.0 or greater. Cytogenetic stocks including F1 hypo-aneuploids, each deficient for a known *G. hirsutum* chromosome, were also genotyped for the same markers.

Phylogenetic Analysis

The VCF file produced from GATK that included genotypes from all 14 samples was imported into R (R Development Core Team 2010). The SNPRelate package (Zheng, Levine et al. 2012) was used to perform a principle component analysis and eigenvalue1 and eigenvalue2 were used to visualize samples. A distance matrix was determined between samples and then used for hierarchical clustering over 10,000 permutations (z.threshold=15, outlier.n=2). The clustering results were used to create a dendrogram tree of the 14 samples.

Results

Candidate SNPs derived from BAC-end sequences

We identified a total of 132,262 intraspecific SNPs for *G. hirsutum* that occurred on average every 888 base pairs (bp). The distribution of base pairs between adjacent intraspecific SNPs found on the same BES is on average 76bp (**Figure 3.1a**). Interspecific SNPs for *G. barbadense* and *G. longicalyx* occurred at a much higher rate than intraspecific SNPs, which was expected due to longer divergence time between the species. A total of 223,138 interspecific SNPs between *G. barbadense* and *G. hirsutum* were determined. While SNPs were similarly spaced (78bp) as the intraspecific SNPs, a larger number of SNPs was identified compared to intraspecific SNPs because they occurred on more BESs (**Figure 3.1b**). A total of 470,631 interspecific SNPs were identified between *G. longicalyx* and *G. hirsutum*. The distance distribution between adjacent SNPs found on the same BES from *G. longicalyx* is quite different from the *G. hirsutum* and *G. barbadense* distributions, and shows that SNPs are more likely to be in close proximity if found on the same BES due to an overall elevated number of SNPs (**Figure 3.1c**). This difference is expected and reflects the greater divergence of *G. longicalyx*. Overlap of SNPs identified in multiple species is shown in **Figure 3.2**. Considering all identified polymorphisms across all three species, a SNP was identified on average every 152 bases and nucleotide diversity across all polymorphic sites or Nei’s Pi was found to be 0.1789.

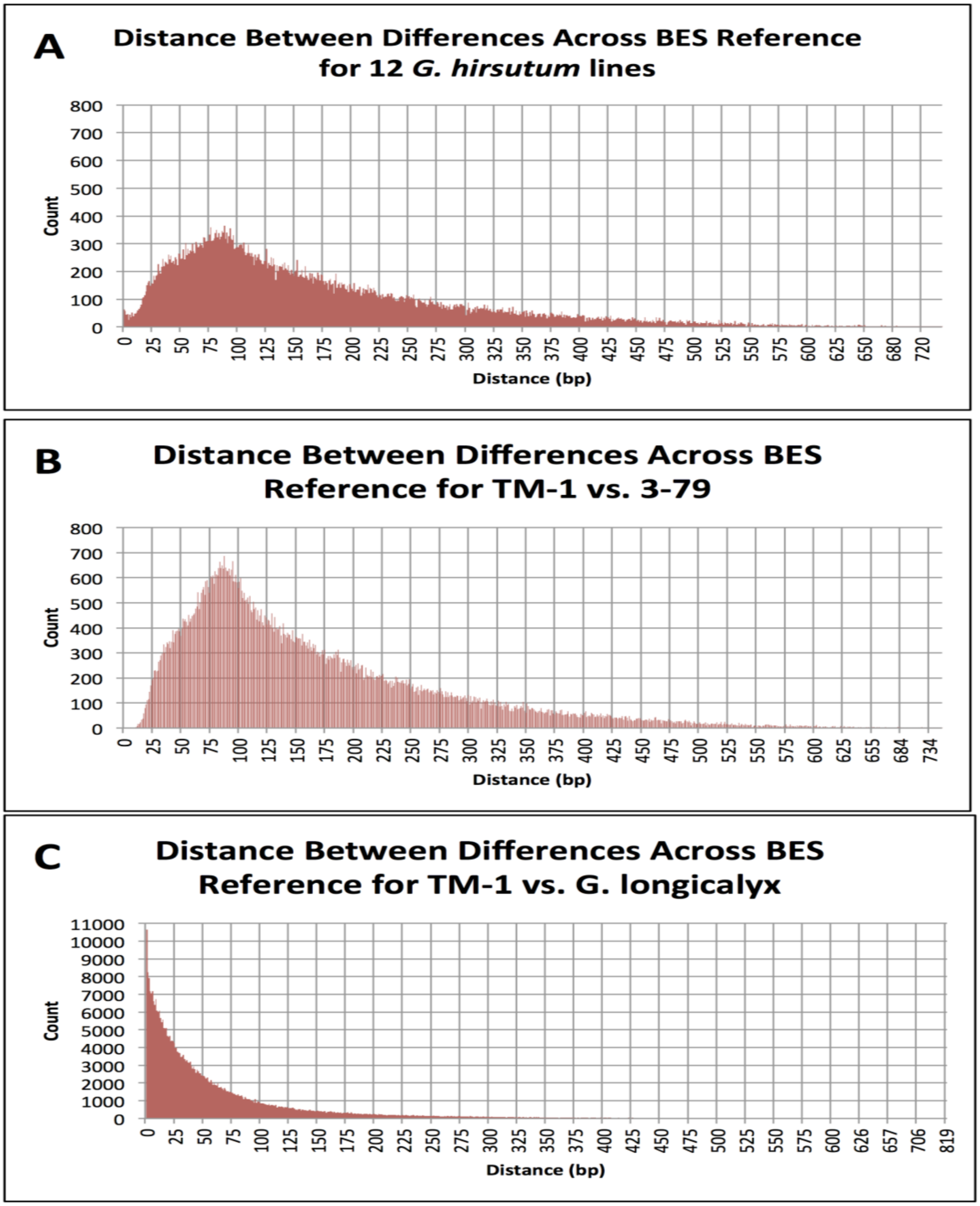


Figure 3.1 Distance between polymorphisms developed in (A.) *G. hirsutum*, (B.) *G. barbadense*, and (C.) *G. longicalyx* relative to the *G. hirsutum*-derived BAC-end sequences.

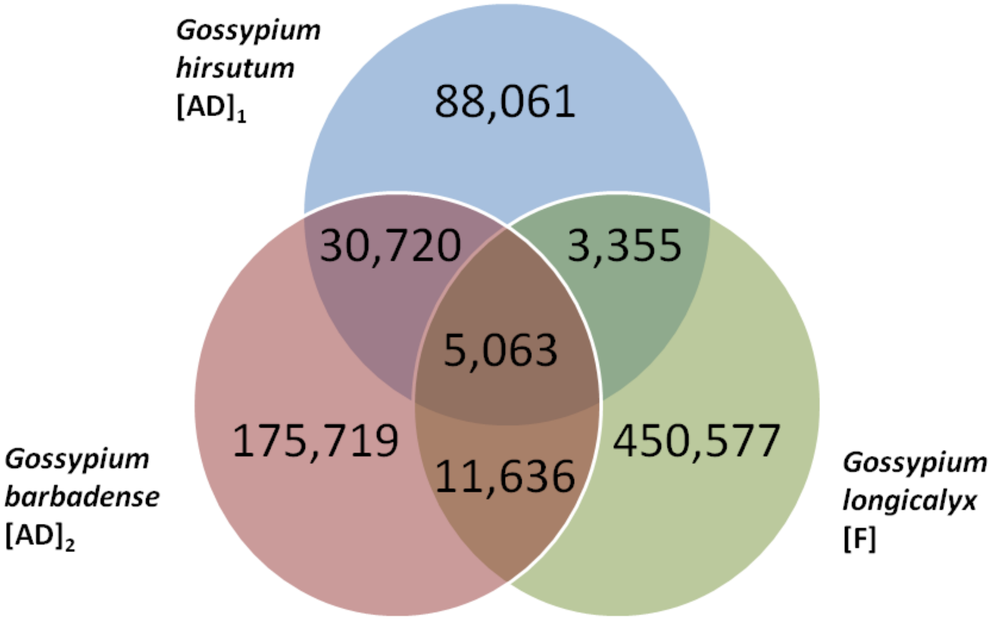


Figure 3.2 Overlap of SNPs identified using *G. hirsutum*, *G. barbadense*, and *G. longicalyx* samples.

Types of SNPs were distributed similarly in all three species, with transitions being more abundant than transversions. However, *G. longicalyx* had a lower transition to transversion ratio (1.63), compared to *G. hirsutum* (2.19) and *G. barbadense* (2.21) (**Table 3.2**). The amounts of missing data and heterozygous loci were also similar between species (**Table 3.3**). The percentage of missing loci was related to evolutionary divergence between the samples analyzed and the reference, with *G. longicalyx*, *G. barbadense* and wild *G. hirsutum* line TX0231 exhibiting the largest percentages of missing data (in decreasing order). High levels of missing data in *G. longicalyx* was expected because as a diploid, it would not contain loci from both A- and D-subgenomes of the tetraploid species, but rather would contain mostly only A-subgenome-like loci being that it is closely related to the A-genome diploids (Phillips and Strickland 1966). Homozygous differences, which are positions for which two samples are both homozygous for a different base at a single locus, between samples were counted for each pair of samples (**Table 3.4**). Overall the number of differences between pairs of cultivated *G. hirsutum* samples was quite variable, ranging from 5,562 (Paymaster HS-26 vs Fibermax 832) to 29,052 (TM-1 vs Tamcot Sphinx). Differences between cultivated samples and the uncultivated *G. hirsutum* TX0231 were in general quite similar, as was the case between all *G. hirsutum* samples with *G. barbadense* and *G. longicalyx*.

Table 3.2 Distribution of SNP types identified in *G. hirsutum*, *G. barbadense*, and *G. longicalyx*.

	G. hirsutum		***G. barbadense***		***G. longicalyx***		**Overall**	
Total SNP	132,262	100.0%	187,355	100.0%	450,577	100.0%	770,194	100.0%
M (A/C)	11,146	8.4%	16,558	8.8%	42,802	9.5%	70,506	9.2%
R (A/G)	45,444	34.4%	64,420	34.4%	139,410	30.9%	249,274	32.4%
W (A/T)	11,992	9.1%	16,021	8.6%	60,070	13.3%	88,083	11.4%
S (C/G)	6,862	5.2%	9,188	4.9%	25,420	5.6%	41,470	5.4%
Y (C/T)	45,408	34.3%	64,574	34.5%	139,620	31.0%	249,602	32.4%
K (G/T)	11,410	8.6%	16,594	8.9%	43,255	9.6%	71,259	9.3%
Total Transition	90,852	68.7%	128,994	68.9%	279,030	61.9%	498,876	64.8%
Total Transversion	41,410	31.3%	58,361	31.1%	171,547	38.1%	271,318	35.2%

Table 3.3 Description of missing data and heterozygous loci in the final VCF for *G. hirsutum*, *G. barbadense*, and *G. longicalyx* samples.

Species	Sample	Genotyping Data			
		Missing (#)	Missing (%)	Heterozygous (#)	Heterozygous (%)
Gossypium hirsutum L.	TM-1	2,144	0.28%	16,880	2.20%
	Sealand 542	4,180	0.54%	22,460	2.93%
	PD-1	4,435	0.58%	27,039	3.53%
	Paymaster HS26	4,450	0.58%	38,201	4.99%
	M-240 RNR	4,146	0.54%	30,450	3.97%
	Fibermax 832	4,994	0.65%	31,862	4.16%
	Coker 312	4,604	0.60%	21,831	2.85%
	SureGrow 747	4,187	0.54%	21,589	2.82%
	Stoneville 474	4,058	0.53%	21,554	2.81%
	Tamcot Sphinx	5,934	0.77%	22,342	2.92%
	Acala Maxxa	5,159	0.67%	20,207	2.64%
	TX0231	12,057	1.57%	24,974	3.29%
Gossypium barbadense L.	3-79	43,530	5.65%	20,443	2.81%
Gossypium longicalyx	F1-1	221,786	28.80%	30,204	5.51%

Table 3.4 Pairwise comparison of homozygous different genotypes between fourteen resequenced samples using BCFtools command gtcheck.

	G. hirsutum												*G. barbadense*	*G. longicalyx*
	TM-1	Sealand542	PD-1	Paymaster HS26	M-240 RNR	Fibermax 832	Coker 312	SureGrow 747	Stoneville 474	Tamcot Sphinx	Acala Maxxa	TX0231	3_79	F1-1
TM-1	-	15,755	20,141	16,177	17,983	18,405	19,418	18,837	16,103	29,052	25,710	88,174	232,669	444,358
Sealand542	15,755	-	11,965	9,062	10,216	11,711	12,283	13,938	11,275	21,305	17,917	94,229	231,726	442,608
PD-1	20,141	11,965	-	9,113	11,672	13,319	13,163	14,435	13,187	20,015	18,488	92,151	229,962	442,294
Paymaster HS26	16,177	9,062	9,113	-	6,937	5,562	7,974	9,948	8,825	13,096	8,821	85,791	224,848	440,842
M-240 RNR	17,983	10,216	11,672	6,937	-	8,865	13,599	12,779	10,929	17,726	16,283	89,815	228,074	442,033
Fibermax 832	18,405	11,711	13,319	5,562	8,865	-	13,462	13,507	11,668	15,290	14,310	89,910	226,029	441,551
Coker 312	19,418	12,283	13,163	7,974	13,599	13,462	-	12,834	11,891	21,663	18,365	94,397	231,818	442,483
SureGrow 747	18,837	13,938	14,435	9,948	12,779	13,507	12,834	-	7,046	21,919	17,913	94,752	232,196	443,104

Table 3.4 (continued)

	G. hirsutum												*G. barbadense*	*G. longicalyx*
	TM-1	Sealand542	PD-1	Paymaster HS26	M-240 RNR	Fibermax 832	Coker 312	SureGrow 747	Stoneville 474	Tamcot Sphinx	Acala Maxxa	TX0231	3_79	F1-1
Stoneville 474	16,103	11,275	13,187	8,825	10,929	11,668	11,891	7,046	-	22,235	17,739	94,936	232,169	442,949
Tamcot Sphinx	29,052	21,305	20,015	13,096	17,726	15,290	21,663	21,919	22,235	-	20,078	98,814	228,004	442,364
Acala Maxxa	25,710	17,917	18,488	8,821	16,283	14,310	18,365	17,913	17,739	20,078	-	94,015	231,328	445,417
TX0231	88,174	94,229	92,151	85,791	89,815	89,910	94,397	94,752	94,936	98,814	94,015	-	230,410	442,605
3_79	232,669	231,726	229,962	224,848	228,074	226,029	231,818	232,196	232,169	228,004	231,328	230,410	-	429,140
F1-1	444,358	442,608	442,294	440,842	442,033	441,551	442,483	443,104	442,949	442,364	445,417	442,605	429,140	-

Conversion of in silico SNPs to assays

G. barbadense SNPs developed *in silico* were randomly, and non-randomly (based on theoretical location on chromosomes 12 and 26 based on alignment to the D_5 sequence), selected for primer design and experimental screening. Initially a random set of 32 *G. barbadense* markers was selected to determine optimal parameters for homeo-SNP removal. Screening produced a total of 13 SNP (40.6%) that were found to have acceptable clustering patterns; of which 7 (53.8%) were associated and 6 (46.2%) were not associated with the transcriptome, respectively (**Table 3.1**). Thus no difference was seen for selecting markers based on association with the transcriptome. When attempting to determine the best parameters for *in silico* SNP calling, results from Iteration-1 and Iteration-2 were compared to determine which iteration resulted in the set of markers with the highest overall validation rate. Taking into account the 13 good markers, Iteration-1 produced a success rate of 61.1% (11/18) while Interation-2 produced a success rate of 84.6% (11/13), therefore Iteration-2 mapping parameters (0.99 length fraction and 0.98 similarity fraction) were utilized for final mapping of all samples, except for *G. longicalyx,* as noted in the Methods (**Table 3.5**).

Utilizing the SNPs identified in *G. barbadense*, following homeo-SNP removal, an additional 96 interspecific SNPs with successfully designed primers were selected for screening based on inferred positioning on chromosomes 12 and 26. A total of 77 (or 80.2%) markers were categorized as good markers.

A total of 48 intraspecific SNP markers were screened on 32 *G. hirsutum* samples and *G. barbadense* line 3-79 which produced 40 good markers or an 83.3%

success rate (**Additional File 3.3**). In order to obtain primers for the 48 markers 87 SNPs were randomly selected for primer design as primer design was successful in 56% of cases. Sample call rates ranged from 0.175 – 1.000. Excluding the two lowest samples all samples had greater than 72.5% call rate. Marker call rates ranged from 0.545 – 0.970.

Table 3.5 Sequencing and mapping statistics for analyzed samples.

Species	Sample	Raw Reads (#)	Trimmed Reads		Mapped Reads		Fraction of Reference Covered
			(#)	(%)	(#)	(%)	
Gossypium hirsutum L.	TM-1	816,832,880	796,773,046	97.5%	59,000,908	7.4%	84%
	Sealand 542	853,279,526	830,824,387	97.4%	62,395,682	7.5%	84%
	PD-1	812,843,166	792,473,570	97.5%	61,009,021	7.7%	84%
	Paymaster HS26	882,495,892	864,048,129	97.9%	67,174,511	7.8%	84%
	M-240 RNR	890,769,668	865,155,108	97.1%	64,913,995	7.5%	84%
	Fibermax 832	804,893,636	790,082,996	98.2%	62,435,893	7.9%	84%
	Coker 312	784,902,894	766,214,459	97.6%	56,052,833	7.3%	84%
	SureGrow 747	897,203,024	871,935,172	97.2%	64,302,689	7.4%	84%
	Stoneville 474	852,597,120	831,199,060	97.5%	59,158,880	7.1%	84%
	Tamcot Sphinx	777,426,896	758,074,043	97.5%	57,448,801	7.6%	84%
	Acala Maxxa	927,373,608	918,133,054	99.0%	67,051,425	7.3%	84%
	TX0231	861,266,926	837,855,860	97.3%	61,276,636	7.3%	83%
Gossypium barbadense L.	3-79	826,774,068	805,308,483	97.4%	49,280,456	6.1%	77%
Gossypium longicalyx	F1-1	1,068,517,678	1,048,843,534	99.0%	64,892,839	6.2%	36%

A total of 32 *G. longicalyx* interspecific SNP markers amenable to successful primer design were selected for screening. Primer design for this diploid species was significantly more difficult compared to the other species due to the elevated number of SNPs and their close proximity within a BES (**Figure 3.1c**). Upon screening, a total of 31 (96.9%) SNPs were determined to produce good assays (**Additional File 3.4**). Two of the markers listed as good however exhibit very close clusters (GH_TBb029G06r260, GH_TBb048K02r139), the proximity of which may lead to difficulties for cluster separation with additional samples. So conservatively, 29 would be regarded as good markers, i.e., a 90.6% success rate.

Alignment of markers to diploid cotton genomes

When all markers from all three species identified were aligned to the high-quality *G. raimondii* (D_5) reference (Paterson, Wendel et al. 2012) genome 38.8% of markers could be aligned to the genome. This was distinctly different from transcriptome derived markers, reported in Hulse-Kemp et al. (2014), of which 75.9% could be aligned to the D_5 genome. The starkly different percentage of mapped markers is likely due to the higher sequence conservation in genic regions. Markers were also aligned to the *G. arboreum* (A_2) draft genome (Li, Fan et al. 2014). The percentages of markers aligning to the A_2 and D_5 reference genomes was very similar for *G. hirsutum* (70.6%/40.9%) and *G. barbadense* (69.6%/41.1%), whereas the percentage was very different for *G. longicalyx* (55%/18.4%). A larger fraction of markers which can be aligned to the A_2 genome is expected for *G. longicalyx* as it is an F-genome diploid that is closely related to A-genome diploids, including *G. arboreum*, and it is relatively distant from the D_5 species, *G. raimondii* (Wendel, Brubaker et al. 2009).

Utilizing the alignment information to both A- and D-diploid reference genomes, markers can theoretically be localized to either the A- or D-subgenome in the tetraploid by considering if the markers align uniquely to either the A_2 or D_5. In both *G. hirsutum* and *G. barbadense*, ~52-53% of markers aligned uniquely to A_2 so can be putatively localized to the A-subgenome, while ~22-24% of markers aligned uniquely to D_5 and can be putatively localized to the D-subgenome. As the A_2 is approximately twice the size of the D_5, it is expected that a larger percentage of markers should be uniquely localized to A_2 (Hendrix and Stewart 2005). *G. longicalyx* had approximately the same proportion of A-specific markers, but the proportion of D-specific markers is considerably lower (12.6%) which is expected being closely related to A-genome diploids.

Interspecific linkage mapping of candidate SNPs and anchoring of contigs

From the random and non-random (selected for chromosomes 12 and 26 based on alignment to D_5 reference sequence) screened SNPs identified from *G. barbadense*, genotypes were obtained for 88 markers that represent 67 unique BAC-end sequences. These 88 markers were screened against a population of 118 F2 (TM-1 x 3-79) individuals and F1 hypo-aneuploid stocks. Upon linkage mapping of the F2s with JoinMap 4.1 software, two linkage groups representing a total of 236.2 cM were obtained (**Figure 3.3**). As expected, due to being randomly selected from the entire data set, all of the markers tested from the randomly selected set (11) were not linked, and were listed as singletons. The resulting two linkage groups were identified as allotetraploid homeologous chromosomes 12 and 26 by both loss of heterozygosity in F1 hypo-aneuploid samples and by alignment information to the D_5 reference genome.

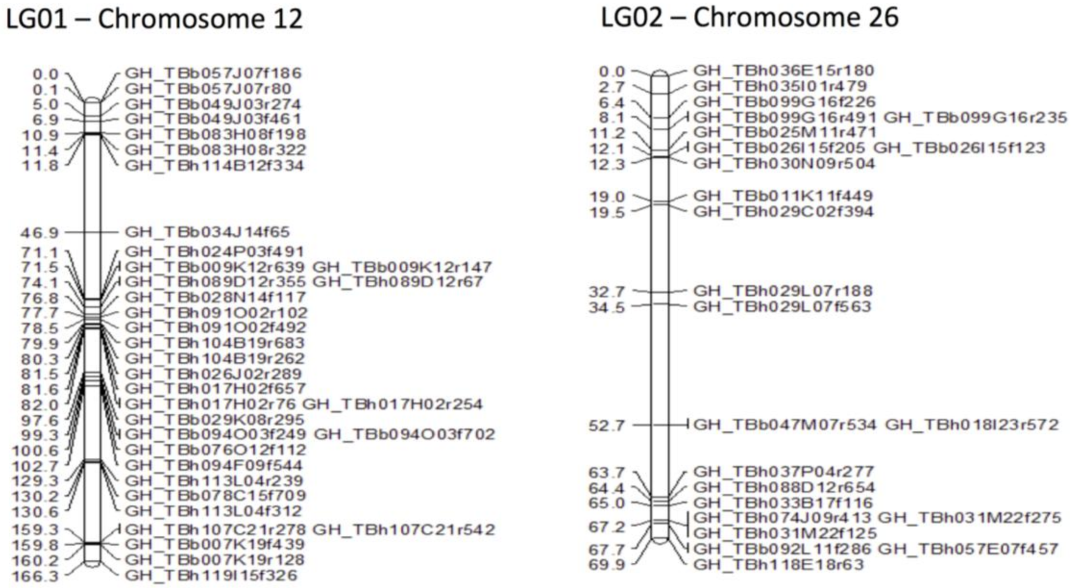


Figure 3.3 Linkage groups determined utilizing 118 interspecific (*G. barbadense* line 3-79 x *G. hirsutum* line TM-1) F_2 samples for BAC-end derived SNPs in JoinMap.

PCA and dendrogram analysis

Principle component analysis with the SNPRelate program was able to successfully separate *G. hirsutum* samples from the other species, *G. barbadense* and *G. longicalyx* (**Figure 3.4**). The analysis also showed a slight difference with cultivated *G. hirsutum* lines and the one wild *G. hirsutum* line TX0231. Similarly a dendrogram compiled with the SNPRelate program (**Figure 3.5**) was able to separate the wild species, and indicated *G. longicalyx* as the out-group, as expected. It also showed the kinship coefficient between *G. hirsutum* samples to be extremely high (near 1) and the individual dissimilarity, or difference between individuals, to be very low (close to 0).

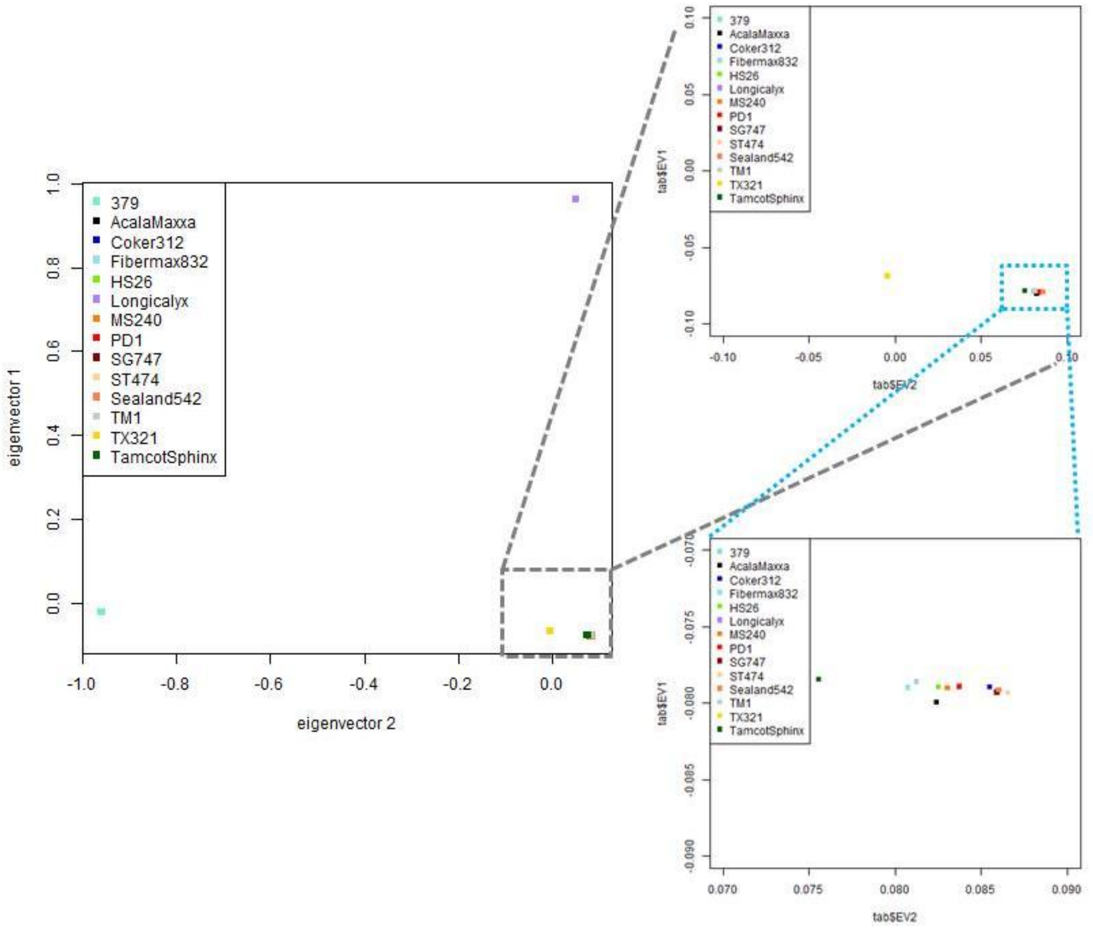


Figure 3.4 Principle component analysis utilizing BAC-end sequence derived SNPs for twelve *G. hirsutum* samples (TM-1, Sealand 542, PD-1, Paymaster HS-26, M-240 RNR, Fibermax 832, Coker 312, SureGrow 747, Stoneville 474, Tamcot Sphinx, Acala Maxxa, TX0231), *G. barbadense* (3-79), and *G. longicalyx* using the SNPRelate package in R.

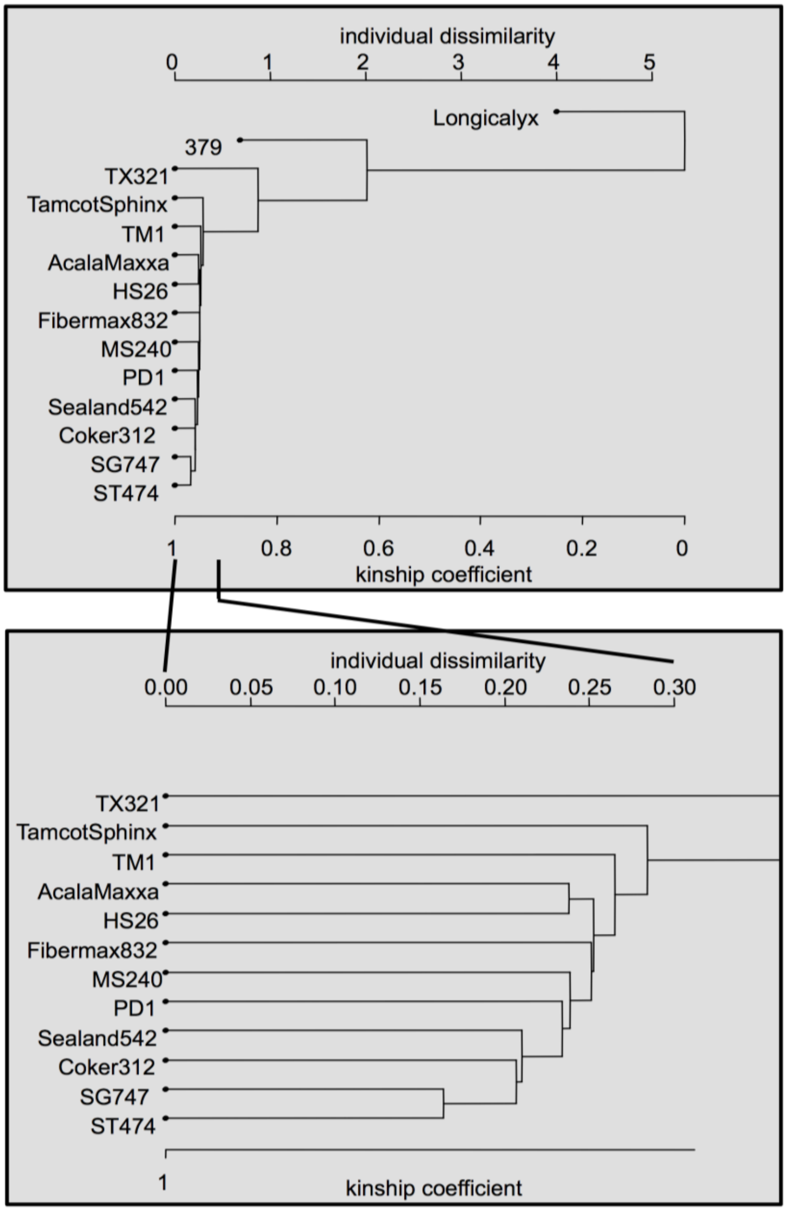


Figure 3.5 Dendrogram produced by hierarchical clustering utilizing BAC-end sequence associated SNPs for 12 *G. hirsutum* samples (TM-1, Sealand 542, PD-1, Paymaster HS-26, M-240 RNR, Fibermax 832, Coker 312, SureGrow 747, Stoneville 474, Tamcot Sphinx, Acala Maxxa, TX0231), *G. barbadense* (3-79), and *G. longicalyx* using the SNPRelate package in R.

Discussion

Reliability of the SNP-based integration of physical and genetic maps

The resulting linkage maps demonstrate the feasibility of integrating physical maps by utilizing markers associated with BAC-end sequences for genetic mapping. In this way a large number of markers can be identified using resequencing data and genetically mapped with a moderately sized mapping population to obtain genetic maps associated with BAC resources. We demonstrate that even with an F2 mapping population of 118 individuals we are able to obtain recombination events within the length of the BAC (average 120 kilobases) and are able to separate markers associated with the forward and reverse reads of a BAC in some instances. Additionally, multiple markers from the same BES were assayed on the population and linkage mapping places all of those markers except for two from GH_TBh104B19r in the same mapping position. The accuracy of mapping BES-associated SNPs makes it possible to utilize SNP mapping for ordering and orienting BACs and BAC contigs when SNPs are identifiable on both forward and reverse end sequences of single BACs, or when multiple SNPs are identified within a given BES.

Factors that affect development of BES-associated SNPs in cotton

Development of genomic-based SNPs in cotton has largely been hindered by availability of a high-quality reference, which has led to development efforts largely targeting transcriptome-based SNPs or SNPs identified using reduced representation libraries. In general, genomic enablement for modern cultivated cotton through marker-assisted breeding is constrained by extremely low diversity and large identical regions between the subgenomes. Because of the high similarity between homeologous chromosomes and particularly genic regions it is extremely difficult to use short next generation sequencing technology derived reads to uniquely localize them to a reference sequence. The availability of high-quality BAC-end sequences from a newly developed BAC library resource has provided a high-quality reference of Sanger reads. With a relatively even distribution across the subgenomes, the BESs are a superior reference with very high quality, allowing mapping of short NGS reads obtained from resequencing that have much higher error rates. In this study, this is achieved by using very high stringencies over the entire length of quality-trimmed NGS reads. However, this mapping approach is not feasible with references that are not specific to allotetraploid cotton. When the reference is known to be of much higher quality compared to typical next-generation assemblies, assay development from *in silico*-derived SNPs using this reference is also greatly enhanced. The flanking sequence based on the BES reference should theoretically be correct for TM-1 which can represent cultivated cotton with expectedly low levels of overall diversity and high levels of synteny; this hypothesis holds up to experimental testing as large number of SNPs are able to be successfully genotyped. This is likely due to the automatic inclusion of correct haplotype information for homeo-SNP alleles which has been shown to increase percentage of co-dominant marker and overall success rate within polyploid cotton (Islam, Thyssen et al. 2014).

While success-rates of the BES-associated SNPs are higher than previous NGS and RRL approaches (Van Deynze, Stoffel et al. 2009; Byers, Harker et al. 2012; Rai, Singh et al. 2013; Gore, Fang et al. 2014; Hulse-Kemp, Ashrafi et al. 2014; Islam, Thyssen et al. 2014; Zhu, Spriggs et al. 2014), all *in silico* SNPs were not verified. This could be due to many different reasons, primarily the BES reference only represents a small portion of the cotton genome, mostly non-coding intergenic regions, and thus while reads are required to have a unique mapping to the reference, it is still possible additional instances of the sequence exists in the overall cotton genome. This is particularly likely due to ancient paleo-polyploid events that have been discovered in the cotton evolutionary lineage (Paterson, Wendel et al. 2012). Even in largely unique genic regions, gene families and duplicates are common, which will create falsely identified SNP that when experimentally assayed will produce unidentifiable clusters and result in unreliable or bad markers.

Utilization of BES-associated markers in G. hirsutum germplasm

The availability of large numbers of intraspecific SNP markers is essential for marker-assisted breeding in *G. hirsutum* lines, particularly with elite cultivars, which exhibit very low levels of polymorphism. Within a typical elite-by-elite single cross, only a very small percentage of markers will subsequently segregate (typically <5%), so it is difficult to map large numbers of markers from intraspecific crosses. To date, this has led to use of interspecific crosses for the majority of mapping in allotetraploid cotton, such as the one utilized in this effort *G. hirsutum* (line TM-1) by *G. barbadense* (line 3-79) (Yu, Kohel et al. 2012). However, in this study we see that even within the small set of 48 experimentally tested intraspecific markers, most of the 32 tested *G. hirsutum* lines can be uniquely identified. Even the three TM-1 samples obtained from different labs had a small number of identified differences. The KASP assays were largely successful for most samples except for Fibermax 966 and Deltapine 90 which likely had a much lower DNA concentration when measured via nanodrop so was not reaching a genotyping end-point consistently with the other samples and thus lead to many uncalled genotypes. Call frequencies for individual markers showed greater variability than sample call frequencies, which is likely due to the need for optimization of PCR conditions for different marker sets. However when using the Fluidigm system, this is difficult as all marker sets are run under the same conditions. It was found that while SNPs were identified in a relatively small number of samples from US lines, polymorphisms were also transferrable to other germplasm sources, such as the three Australian samples (Delta Opal, Sicot 70, and Siokra 1-4) and one Indian line (MCU-5) included on the panel. This indicates that ascertainment bias may not be very large, and that sample lineage is likely more important than the country of germplasm origin.

While only small differences exist between lines, the PCA and hierarchical clustering analyses were able to distinguish relationships among the samples. In the PCA, the wild *G. hirsutum* line TX0231 was markedly different from the cultivated samples, and the Tamcot Sphinx (a MAR program derived sample) was the next most unique sample, which was expected due to the assumed diversity of those samples. Within the cultivated lines, when zoomed in to a very small section of the plot, the samples appear to occupy three different clusters. Cluster 1 contains TM-1 and Fibermax 832, cluster 2 contains Acala Maxxa, Paymaster HS-26, M-240 RNR, and PD-1 and cluster 3 contains Sealand 542, SureGrow 747, Stoneville 474 and Coker 312. In Fang et al. (2013), three of the included samples, SureGrow 747, Stoneville 474 and Fibermax 832, were all identified as being in the same group (Group 6). While we found two of the samples in the same cluster, Fibermax 832 was found in a separate cluster. Fang et al. also identified Acala Maxxa and Paymaster HS-26 to occur in the same group (Group 7), which correlates with our analysis where both are in cluster 2. The PCA results correlate with the hierarchical clustering analysis with the shared samples showing the greatest relationship: Acala Maxxa and Paymaster HS-26 from Fang Group 7; SureGrow 747 and Stoneville 474 from Fang Group 6. The TM-1 sample, which was the sample utilized as a reference, was also found to have the largest individual dissimilarity after the MAR sample (Tamcot Sphinx). It may be possible that this information is correct as TM-1 was a line established in mid-twentieth century (Kohel, Richmond et al. 1970), however it is also likely that due to ascertainment bias from using this sample as the reference that it looks more different to the rest of the other cultivated *G. hirsutum* lines.

Development of BES-associated markers is helpful for integration of physical and genetic maps

The abundance of markers and accuracy of localization using SNPs associated with BAC-end sequences will be extremely helpful for integrating an allotetraploid cotton physical map with genetic maps. Upon finalization of an allotetraploid cotton physical map, BES-associated SNPs can also be utilized to integrate unplaced contigs and singletons to enhance the completion of a quality draft reference sequence. Fine-mapping utilizing BES-associated SNPs with a large population can also be used to correct ordering and localization of contigs. Integrating genetic maps with quantitative trait loci (QTL) mapping will allow for utilization of BAC resources for QTL cloning and fine-scale investigation of important regions in the cotton genome.

CHAPTER IV

DEVELOPMENT OF A 63K SNP ARRAY FOR *GOSSYPIUM* AND HIGH-DENSITY MAPPING OF INTRA- AND INTER-SPECIFIC POPULATIONS OF COTTON (*GOSSYPIUM* SPP.)

Introduction

Cotton (*Gossypium* spp.) is the world's most important renewable natural textile fiber crop and also a significant source of oilseed. Cotton is grown in over 75 countries and produced over 118 million bales of fiber world-wide in 2013 (National Cotton Council, www.cotton.org). In the United States, the 2013 crop of 12.9 million bales was valued at $5.2 billion and had an estimated overall direct economic impact of $27.6 billion (US Department of Agriculture; National Agriculture Statistics Service, www.nass.usda.gov). Most of the cotton fiber is used for apparel products (75%), and small percentages are used for home furnishings (18%) and industrial products (7%) (National Cotton Council, www.cotton.org). In the past, cottonseed has been regarded mostly as a byproduct of the crop, but has recently become marketable as a protein source for livestock, dairy cattle, and poultry, as well as for production of cottonseed oil, which is used in the food product industry. Contemporary cotton production relies primarily on allotetraploid *Gossypium hirsutum* L. (2*n*=4*x*=52), that contributes over 95% of the world crop. The remaining production relies largely on *G. barbadense* L., another allotetraploid, and to lesser extents on two A-genome diploids (2*n*=2*x*=26), *G. arboreum* L. and *G. herbaceum* L., primarily in Asia. The *Gossypium* genus contains approximately 51 species, including 6 allotetraploid species and 45 diploid species (Wendel, Brubaker et al. 2009; Ulloa, Abdurakhmonov et al. 2013; Grover, Zhu et al. 2014). While diploid species of interest such as *G. longicalyx*, an F-genome diploid, and *G. armourianum*, a D-genome diploid, cannot be readily crossed with the primary cultivated species *G. hirsutum*, the allotetraploid species such as *G. barbadense*, *G. tomentosum,* and *G. mustelinum* can be readily intercrossed.

Allotetraploid cotton was formed approximately 1-2 million years ago (mya) (Wendel, Brubaker et al. 2009) in a polyploidization event between two diploids: one with an A-like genome and the other with a D-like genome. Thus, the ancestral genomes would resemble the genomes of extant diploids *G. arboreum* (A_2) and *G. raimondii* (D_5). The *G. hirsutum* genome is designated $[AD]_1$, and has a C-value of approximately 2.4 Gbp; the 26 chromosomes are numbered according to genome ancestry, where chromosomes 1-13 are of A-genome origin and 14-26 are of D-genome origin (Brown 1980; Wendel, Brubaker et al. 2009). Because the ancestral genomes of cotton diverged only 5-10 mya, and the polyploidization event was 1-2 mya, initial efforts to develop single nucleotide polymorphism (SNP) markers were hindered by the co-identification of interlocus SNP variants between the two subgenomes in the tetraploid, or homeo-SNPs. Recent developments from the cotton community including the publication of a high-quality genome reference sequence for *G. raimondii* (Paterson, Wendel et al. 2012) and draft sequences for *G. arboreum* (Li, Fan et al. 2014) and *G. raimondii* (Wang, Wang et al. 2012) have aided development of large SNP data sets in gene-based and genome-based identification efforts (Van Deynze, Stoffel et al. 2009; Byers, Harker et al. 2012; Lacape, Claverie et al. 2012; Rai, Singh et al. 2013; Gore, Fang et al. 2014; Hulse-Kemp, Ashrafi et al. 2014; Islam, Thyssen et al. 2014; Zhu, Spriggs et al. 2014). Increasingly larger genetic maps with greater resolution and saturation have been generated as numbers of primarily SSR markers increased, culminating in the recent publication of a consensus genetic map that comprises more than 8,200 loci based on primarily six inter-specific (*G. hirsutum* x *G. barbadense*) populations (Blenda, Fang et al. 2012). Few of the genetic maps have contained SNP markers, and those have been primarily limited to inclusion of a couple hundred SNPs with SSRs (Byers, Harker et al. 2012; Yu, Kohel et al. 2012; Gore, Fang et al. 2014) to at most a couple of thousand (Islam, Thyssen et al. 2014; Zhu, Spriggs et al. 2014). SNPs have been difficult to develop *in silico*, due to low polymorphism and low divergence among polyploid genomes (Van Deynze, Stoffel et al. 2009). The increasing efficiency of next generation sequencing and improved *in silico* methods have allowed SNP development at the whole genome level even for the 2.4 Gbp allotetraploid cotton genome (Hulse-Kemp et al. 2014).

In cultivated cotton, as well as in other crop species, there is considerable interest in being able to genotype a large number of SNP markers in a high-throughput manner. With advancing array technology, large-scale genotyping can be performed in a massively parallel fashion to assay thousands of loci simultaneously in a short time. Genotype data obtained from the high-throughput assay can then be utilized to evaluate genetic diversity, construct genetic linkage maps, dissect the genetic architecture of important traits (Truco, Ashrafi et al. 2013) and in many more novel applications (Ganal, Polley et al. 2012). SNP arrays for many crop species have recently been developed and have been utilized to advance breeding and discoveries in those crop systems (Ganal, Durstewitz et al. 2011; Hamilton, Hansey et al. 2011; Chen, Xie et al. 2013; Song, Hyten et al. 2013; Truco, Ashrafi et al. 2013; Bianco, Cestaro et al. 2014; Dalton-Morgan, Hayward et al. 2014; Tinker, Chao et al. 2014; Wang, Wong et al. 2014).

The objectives of this study were 1) to utilize recently identified SNPs to develop a standardized large-scale SNP genotyping array for cotton, 2) to evaluate the performance and reproducibility of the array on a large set of samples, 3) to develop a cluster file which can be used to help automate genotyping for allotetraploid cotton, and 4) to produce high-density linkage maps based on two biparental F2 populations in an intra-specific cross (*G. hirsutum* x *G. hirsutum*) and an inter-specific cross (*G. hirsutum* x *G. barbadense*).

Materials and Methods

Plant materials

Seeds for each line were planted in peat pellets (Jiffy, Canada) at Texas A&M University, CIRAD/EMBRAPA, USDA-ARS-SPARC, USDA-ARS-SRRC, or CSIRO greenhouses. Plants were allowed to grow until first true leaves were available. Young true leaves were sampled and extracted according to the manufacturer’s instructions using the Macherey-Nagel Plant Nucleo-spin kit (Pennsylvania). All DNA samples were quantified using PicoGreen® and then diluted to 50ng/µl. The samples included reference lines from the various SNP development efforts, duplicated DNA samples, individual plants and plant pools from the same seed source, individual plant samples from different seed sources, parent/F_1 combinations, segregating samples from wild *Gossypium* species, inbred cultivar lines, wild *G. hirsutum* lines and two mapping populations, one intra-specific and one inter-specific (**Table 4.1**). This material represented samples from most of the cross-compatible range of *G. hirsutum*. Mapping samples included 93 F_2 lines from a *G. hirsutum* cv. Phytogen 72 x *G. hirsutum* cv. Stoneville 474 population designated as "PS" for intra-specific mapping and 118 F_2 lines from a *G. hirsutum* cv. Texas Marker-1 x *G. barbadense* doubled haploid line 3-79 population designated as "T3" for inter-specific mapping. Samples for the mapping populations were chosen randomly from germinated seed within each population.

Array design

SNP data sets were obtained for nine intra-specific (**Table 4.2**) and four inter-specific SNP development efforts (**Table 4.3**). The datasets obtained included SNPs from: [1] restriction enzyme double-digest procedure in *G. hirsutum* between a cultivated and wild accession (Byers, Harker et al. 2012), [2] long-read 454 hypomethylated restriction-based genomic enrichment sequencing of six *G. hirsutum* lines (Rai, Singh et al. 2013), [3] genotyping-by-sequencing (GBS) mapping of a cross between two *G. hirsutum* lines (Gore, Fang et al. 2014), [4] transcriptome sequencing of five *G. hirsutum* lines (Ashrafi et al. – In review), [5] GBS mapping of a multi-parent population (Islam, Thyssen et al. 2014), [6] transcriptome sequencing and restriction-based approach of 18 *G. hirsutum* lines (Zhu, Spriggs et al. 2014), [7&8] re-sequencing of 12 *G. hirsutum* lines (Hulse-Kemp et al. – in review and Ashrafi et al. – Personal Communication), [9] unclassified markers mapped to known chromosomes (unpublished, DOW AgroSciences), [10] single-copy sequence derived SNPs between *G. hirsutum* and *G. barbadense* (Van Deynze, Stoffel et al. 2009), [11] differential expression analysis using RNA-sequencing of *G. hirsutum* versus *G. barbadense* (Lacape, Claverie et al. 2012), [12] transcriptome sequencing of *G. barbadense*, *G. tomentosum*, *G. mustelinum*, *G. armourianum*, and *G. longicalyx* (Hulse-Kemp et al. 2014), and [13] re-sequencing of *G. barbadense* (Hulse-Kemp et al. – submitted).

Table 4.1 Samples included for array validation and cluster file development.

Sample Type	Number of Samples
Inbred - *G. hirsutum* (Cultivated)	516
Inbred - *G. hirsutum* (Wild)	59
Intraspecific F_1	53
Intraspecific F_2	157
Intraspecific Backcross	31
Intraspecific RIL	34
Inbred - *G. barbadense*	18
Interspecific F_2	69 (49*)
Interspecific RIL	14
Interspecific Aneuploid	21**
Wild Tetraploid Species	4
Synthetic Tetraploid	3
Diploid Species	8**
Interspecific Backcross	146
Interspecific F_1	20
Haploid	3**
Total	**1156**

*A total of 49 inter-specific F_2 samples were not included in cluster file development, but were genotyped using the resulting cluster file for inclusion in linkage mapping. ** These samples were used in the cluster file development but the cluster file is not suitable for scoring such samples since it is only optimized for tetraploid samples.

Table 4.2 Datasets utilized in intra-specific content design on the CottonSNP63K array. A total of 50K putative single nucleotide markers were used to produce the 45,104 intra-specific assays on the array after production.

Data Set Name	Authors/Ref	Lines
Brigham Young University	Byers et al. (2012)	Acala Maxxa, TX2094
CSIR-NBRI	Rai et al. (2013)	JKC703, JKC725, JKC737, JKC770, MCU-5, LRA5166
USDA - Set 1	Gore et al. (2014)	TM-1, NM24016
UC-Davis/TAMU GH RNA-seq	Ashrafi et al. (submitted - 2014)	TM-1, FM832, Sealand 542, PD-1, Acala Maxxa
USDA – Set 2	Islam et al. (2014)	Acala Ultima, Pyramid, Coker 315, STV825, FM966, M-240 RNR, HS26, DP-90, SG747, PSC355, STV474
CSIRO	Zhu et al. (2014)	MCU-5, Delta Opal, Sicot 70, Siokra 1-4, DP-16, Tamcot SP37, Namcala, Riverina Poplar, LuMein 14, Sicala 3-2, Sicala 40, Sicala V-2, Sicot 81, Sicot 71, Sicot 189, Sicot F-1, Deltapine 90, Coker 315
TAMU/UC-Davis Intra Genomic – Set 1	Hulse-Kemp et al. (unpublished)	M-240 RNR, TM-1, HS26, SG747, STV747, FM832, Sealand542, PD-1, Coker 312, Tamcot Sphinx, TX231, Acala Maxxa
UC-Davis/TAMU Intra Genomic – Set 2	Ashrafi et al. (unpublished)	M-240 RNR, TM-1, HS26, SG747, STV747, FM832, Sealand542, PD-1, Coker 312, Tamcot Sphinx, TX231, Acala Maxxa
DOW AgroSciences	DAS (unpublished)	Unreleased

Table 4.3 Datasets utilized in inter-specific content design on the CottonSNP63K array. A total of 20K putative single nucleotide markers were used to produce the 17,954 inter-specific assays on the array after production.

Data Set Name	Authors/Ref	Lines
UC-Davis Inter	Van Deynze et al. (2009)	*G. barbadense* (3-79), *G. hirsutum* (TM-1)
CIRAD	Lacape et al. (2012)	*G. barbadense* (VH8-4602), *G. hirsutum* (Guazuncho II)
TAMU/UC-Davis Inter RNA-seq	Hulse-Kemp et al. (2014)	*G. barbadense* (3-79), *G. tomentosum, G. mustelinum, G. armourianum, G. longicalyx*
TAMU/UC-Davis Inter Genomic	Hulse-Kemp et al. (unpublished)	*G. barbadense* (3-79)

The Illumina Infinium technology utilizes a bead-based approach where 50 base pair oligonucleotide probes are used to hybridize to SNP-adjacent sequences of a sample and then a single base pair extension is performed to assay the SNP base with fluorescently labeled nucleotides. The technology utilizes a two-fluorophore system, necessitating the use of two beadtypes to discriminate only those SNPs that target alleles sharing the same fluorophore (transversions) and only one beadtype to capture all other SNP types (transitions). Based on the Infinium technology, putative markers were filtered to maximize value from this technology. Putative SNPs were filtered by design score >0.8 (Illumina, Inc.), identity match >99% within 100bp of flanking sequence, >99% identity match for designed probe sequence, with final SNPs retained based upon precedent publication order in the hierarchy listed above. Subsequent filtering steps included prioritization for 1) one beadtype (Infinium II) assays, 2) validated markers, 3) experimental screening success rates, and 4) representation of genic and non-genic SNPs to optimally cover the cotton genome, to obtain a final set of 70,000 putative SNP markers for inclusion on the array (**Additional File 4.1**).

Due to limited prior validation, 500 markers were randomly selected for inclusion from the UC-Davis/TAMU Genomic – Set 2 and from the CSIR-NBRI data set, 252 were chosen randomly and 134 markers were selected for inclusion based on overlap as identified with BLAST to have 100% identity with markers from other data sets over the entire length of SNP and flanking sequences. For USDA-Set 2, markers were primarily included for contigs that had low missing data and were detected across multiple lines with minor allele frequency greater than 0.1. The markers chosen from the CIRAD set primarily represent markers from the most highly differentially expressed transcripts identified between *G. hirsutum* and *G. barbadense*. From the TAMU/UC-Davis Inter RNA-Seq data set, as many as possible common SNPs between the five species represented and *G. hirsutum* were selected for inclusion, as well as sets of private SNPs unique to each of the five species (**Figure 4.1**).

Genotyping with the array

Standardized DNA at 50ng/µl for each of the cotton lines described above was processed according to Illumina protocols and hybridized to the CottonSNP63K array at Texas A&M University or CSIRO. Single-base extension was performed and the chips were scanned using the Illumina iScan. Image files were saved for cluster file analysis. All image files were uploaded into a single GenomeStudio project containing 1,156 individual samples. Of the 70,000 SNPs targeted for manufacture on the array, 6,942 markers failed to meet standards for bead representation and decoding metrics during the array construction process at Illumina and were removed from the manifest. Data for the remaining markers were clustered using the GenomeStudio Genotyping Module (V 1.9.4, Illumina, Inc.). All markers were viewed and manually curated, taking into account the sample type and known segregation ratios, for construction of the best cluster file for genotyping tetraploid cotton (available at: http://www.cottongen.org/node/add/cotton-cluster-file-request).

Reproducibility and call rate in a diversity set of cotton samples

Three technical replicates were processed for an individual DNA sample of *G. hirsutum* cultivar TM-1, *G. barbadense* inbred line 3-79, and for the F_1 individual from a *G. hirsutum* (TM-1) by *G. barbadense* (3-79) cross. Different individuals from the same line were analyzed for 11 lines from the same seed source and 11 lines from different seed sources. Genotypes from these lines were compared to determine similarity across technical replicates, different seed sources and the same seed source. Multiple individuals (12) were pooled into the same DNA sample and then compared to the genotype from a single individual to determine variability within a seed source as well as percent residual heterozygosity. Polymorphic SNPs were classified based on Illumina

GenTrain score (proximity of clusters) and call frequencies across samples. Minor allele frequencies of polymorphic markers were determined using only inbred line samples.

Genetic linkage analysis – intra-specific and inter-specific

Genotyping data were transformed into mapping data format ("ABH") for the 93 intra-specific PS F2 samples and the 118 inter-specific T3 F2 samples. Only markers that had opposite homozygous allele calls between parental samples and behaved co-dominantly were retained. Subsequently the data files were initially mapped with JoinMap 4.0 (Van Ooijen 2006) with respect to verification of segregation patterns, the formation of linkage groups and the preliminary position of the markers on the chromosomes using the default grouping settings and the maximum likelihood mapping algorithm to determine groups.

The genetic distances between the markers and their final map position were then manually curated in order to remove problematic markers, regarding the number of crossovers (to limit double crossovers) and the length of the linkage group (to minimize total length) using the ABH mapping data file in Microsoft Excel. The final linkage groups were constructed with MapManager QTX (Manly, Cudmore et al. 2001) with the following settings: linkage evaluation F_2, search linkage criterion P=0.05, map function Kosambi, cross type – line cross. Markers were eliminated from the final mapping if they caused unexpected expansion of the linkage group because of too many crossovers since this is likely the result of low scoring quality of the individual marker analysis, i.e. markers with low GenTrain score and/or call frequency. The final maps were drawn with

MapChart version 2.2 (Voorrips 2002) with one marker per centimorgan (cM) to allow easier visualization, positions for all markers are listed in **Additional File 4.1**. Linkage disequilibrium and visualization of crossovers for each determined linkage group were plotted using the CheckMatrix software (www.atgc.org/XLinkage/). The correlation between the inter-specific and intra-specific map orders was visualized using MapChart. Discordant linkage groups in the inter-specific map identified using CheckMatrix and MapChart were remapped in MapManager using a framework map from the corresponding intra-specific linkage group. The number of recombination events per each F_2 individual was calculated for both the intra-specific and the reordered inter-specific linkage maps. Average numbers of recombination bins across both linkage maps were also calculated. Recombination bins were considered as the interval between one recombination breakpoint and the next breakpoint within the mapping populations.

Synteny analyses

All SNP marker sequences were aligned to the high-quality JGI *G. raimondii* (D_5) reference genome (Paterson, Wendel et al. 2012) using Burrows Wheeler Alignment (BWA) software in GALAXY (Goecks, Nekrutenko et al. 2010) with default parameters. Linkage map positions were plotted against D_5 alignment position for both the inter-specific and intra-specific mapped markers. The corresponding allotetraploid chromosomes of the linkage groups were identified using mapped markers (Dow AgroSciences - unpublished; Yu et al. 2014; Blenda et al. 2012) and D_5 alignment information. Linkage groups were oriented using the cotton consensus map which incorporated SSR marker chromosome assignment information available (Blenda, Fang et al. 2012). All markers were also aligned to the BGI *G. arboreum* (A_2) draft sequence (Li, Fan et al. 2014) and plotted.

Results

Genotyping array content

An Illumina Infinium genotyping array was developed targeting 70,000 putative SNP markers (**Additional File 4.1**). The 70,000 putative markers represent 50,000 SNPs that have been identified for use in intra-specific crosses between *G. hirsutum* lines and 20,000 SNPs for use in inter-specific crosses between *G. hirsutum* and other *Gossypium* species, such as *G. barbadense*, *G. tomentosum*, *G. mustelinum*, *G. armourianum*, and *G. longicalyx*. The intra-specific set was developed from an initial set of 1,658,397 putative SNP markers; however, the largest data set (UC-Davis/TAMU Intra Genomic - Set 2) had not been experimentally validated previously and another large data set (CSIR-NBRI) had limited validation, so only small random samples of SNPs for these sets were included. Therefore, selection of the intra-specific markers for the array was based primarily on 91,953 markers. All markers were submitted through the Illumina Design Tool to determine assay design scores for each marker. The data set was filtered to retain only Infinium II, or one-bead type assays that assay one SNP per bead, and those with design scores above 0.8. This resulted in 69,306 SNPs or 75.4% retention. Results of probe design for each of the data sets are shown in **Additional File 4.2**. Duplicated SNPs were eliminated, as noted in the Methods. All available SNP markers

(18,348) within genes (genic) and a randomly selected set of genomic sequence-derived markers (31,396) were included, for a total of 50,000 intra-specific markers. The resulting intra-specific set comprises 36.7% genic SNPs, 62.8% non-genic SNPs, and 0.5% unclassified SNPs.

The same selection and filtering method was used for the inter-specific data set. The SNP selection process started with 314,894 potential markers. After the initial step of filtering based on assay design score and one-bead type markers, there were 234,370 markers. As many previously mapped markers as possible were included. When feasible, a single marker was selected per reference sequence or gene sequence. Additional markers were randomly selected from the filtered data set to fill out the final set of 20,000 inter-specific assays. A total of 14,689 markers in genes were included and 5,311 genomic derived markers were included. The inter-specific final set comprises 73.5% genic SNPs and 26.5% non-genic markers. Within the inter-specific set, the TAMU/UC-Davis Inter RNA-seq markers were selected to contain the maximum number of markers with polymorphism across multiple species (relative to *G. hirsutum*) as possible and approximately 2,000 markers unique to each of the wild species. This allowed for a maximum number of markers that can be used for introgression breeding efforts, while occupying as few beads as possible for cost purposes (**Figure 4.1**). Approximately 18% of the inter-specific set is composed of markers chosen to support work for multiple species germplasm introgression into *G. hirsutum*.

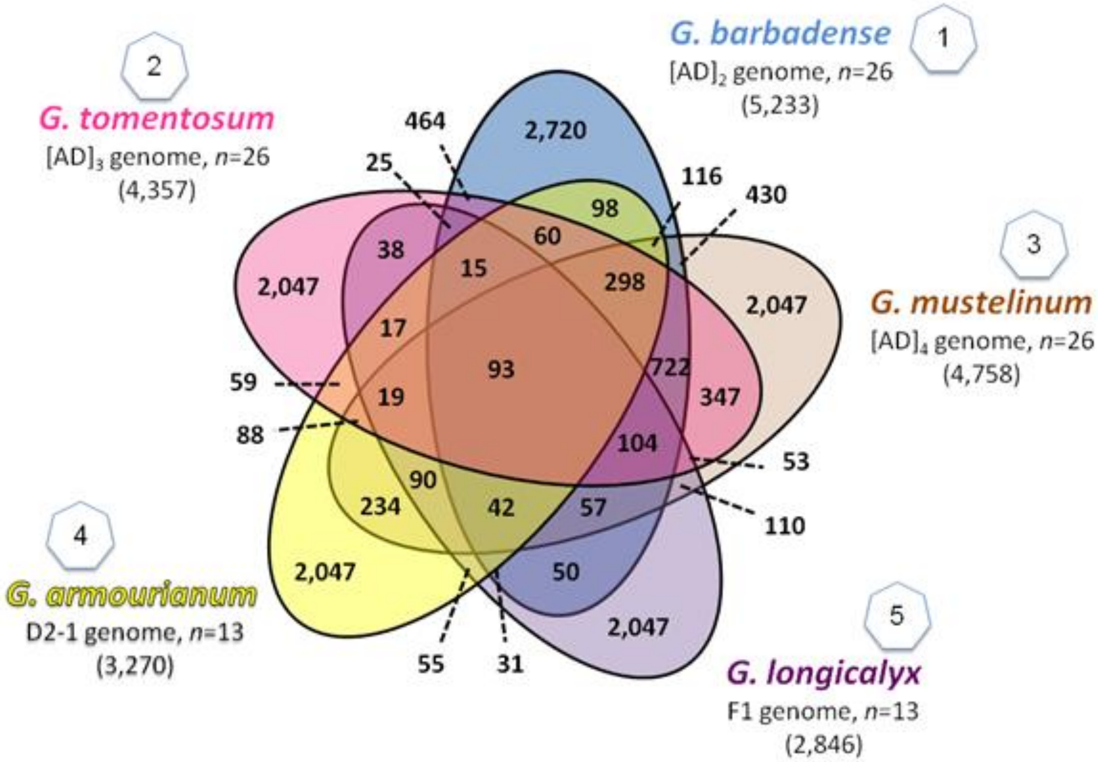


Figure 4.1 SNP markers shared across five species included on the CottonSNP63K array from TAMU/UC-Davis Inter RNA-seq discovery set (Hulse-Kemp et al. 2014).

Together, the selected intra-specific and inter-specific sets resulted in 70,000 SNP markers submitted to manufacturing for inclusion on the array. All these SNPs have been deposited in the CottonGen database (Yu, Jung et al. 2014 www.cottongen.org). Based on the characteristics of the utilized Illumina array format, a total of 90,000 beads can be assayed per sample; therefore it is possible to add up to 20,000 assays, which can be included as private add-on content to allow for adaptability of the array.

Automated genotype calling through cluster definition

From the 70,000 markers sent into production, 63,058 markers passed Illumina manufacturing and quality control, and were analyzed on 1,156 samples for amenability to automated genotyping. Detailed analyses of the diversity and population structure represented by the germplasm used in this project are underway and will be reported separately. The proportion of samples amenable to genotyping, or "call frequency", across all markers on the chip could be grouped into four distinct types (**Figure 4.2a-d**) based on all 1,156 samples that were included for cluster file development were examined. Importantly the polymorphic markers have an average call frequency of 0.99. The first type of classification represents failed markers that did not amplify in the majority of samples (**Figure 4.2a**). The second type consists of markers that had many samples with uncalled genotypes and therefore had a call frequency between 0.50-0.99 (**Figure 4.2b**), which may be caused by the presence of an additional null allele. Type 3 includes markers in which only a few samples remain uncalled (**Figure 4.2c**). The last type is those markers in which all samples are called and are highly reproducible (**Figure 4.2d**). Call frequency distribution of all synthesized SNP markers on the chip are shown in **Figure 4.2e**.

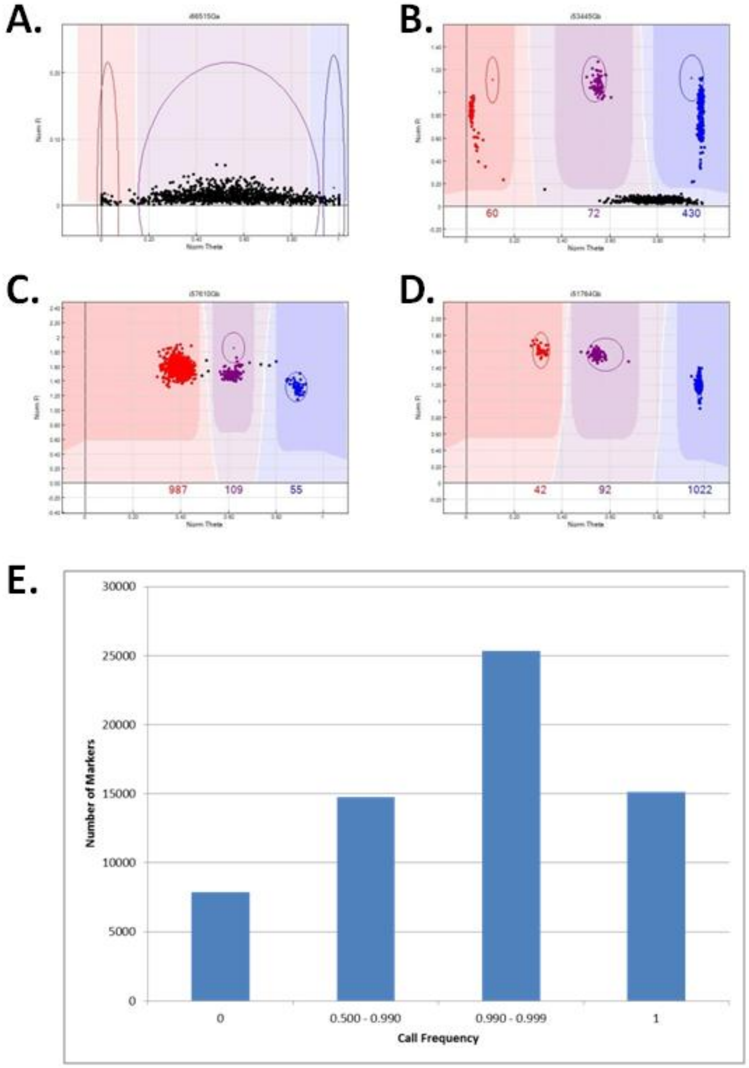


Figure 4.2 Types of call frequency of SNP markers. NormTheta or relative amount of each of the two fluorophore signals is plotted on the X-axis, while NormR or signal intensity is plotted on the Y-axis. **A.**) Failed marker with call frequency = 0, **B.**) Call frequency 0.500-0.990 with major sample deviations, **C.**) Call frequency 0.990-0.999 with few uncalled samples, **D.**) Call frequency = 1 with all called samples, **E.**) Distribution of call frequencies for all SNP markers on the array.

Successful markers primarily produced six distinct clustering patterns (**Figure 4.3a-f**). In the first pattern type essentially all samples fall in a single cluster. This type represents probe sequences that detect a monomorphic locus or loci (**Figure 4.3a**). The second pattern type represents markers which are detecting two monomorphic loci which are each homozygous for a different allele (**Figure 4.3b**). These appear to be heterozygous in all lines are intergenomic or “homeo-SNPs”, that are differences between the two subgenomes of cotton and are often identified as false positives in SNP discovery. Polymorphic markers are attributed to the remaining four patterns; they can be classified according to their GenTrain score. Pattern type 3 are markers which showed three clearly definable clusters and behaved in a traditional co-dominant, diploid-like marker with three possible genotypes (AA, AB, BB) and homozygous genotype clusters located near 0 and 1 (**Figure 4.3c**). These types of markers did not require any significant manual adjustment of marker positions. Similarly, the fourth pattern type (**Figure 4.3d**) showed three clearly identifiable clusters, but the clusters were shifted towards one side of the plot with one homozygous cluster at 0.5 on the X-axis. Pattern 4 markers represent markers that detect two loci, most likely from the homeologous chromosomes: one polymorphic and one monomorphic resulting in three possible genotypes (AAAA, AAAB, AABB). Pattern 5 markers represent a pattern in which the three clusters are quite close together and likely represent three or more loci with one polymorphic locus and multiple monomorphic loci in the background (**Figure 4.3e**). Lastly pattern 6 markers represent markers which show extremely close clusters due to assaying a large number of loci (**Figure 4.3f**). Marker patterns 4 and 5 typically required manual adjustment of locus positions, while pattern 6 markers were frequently set as failed.

After evaluating and manually curating all markers, a total of 55,201 of the markers produced successful assays. The successful assays were further characterized into 38,822 polymorphic markers, 10,314 monomorphic markers and 6,065 intergenomic markers. Thus the overall success rate of the chip is represented by a minimum of 38,822 polymorphic markers (in the samples assayed) out of the 63,058 markers that were synthesized on the chip (61.6%) that will be useful for genotyping cotton samples. Analysis of minor allele frequency (MAF) across polymorphic markers using only inbred samples showed that 66.8% of the polymorphic markers have a MAF above 0.05, 55.8% above 0.10, and 40.0% above 0.2. The average MAF among polymorphic markers on the CottonSNP63K array was found to be 0.17. Distribution of minor allele frequencies of polymorphic markers in inbred samples is shown in **Figure 4.4**.

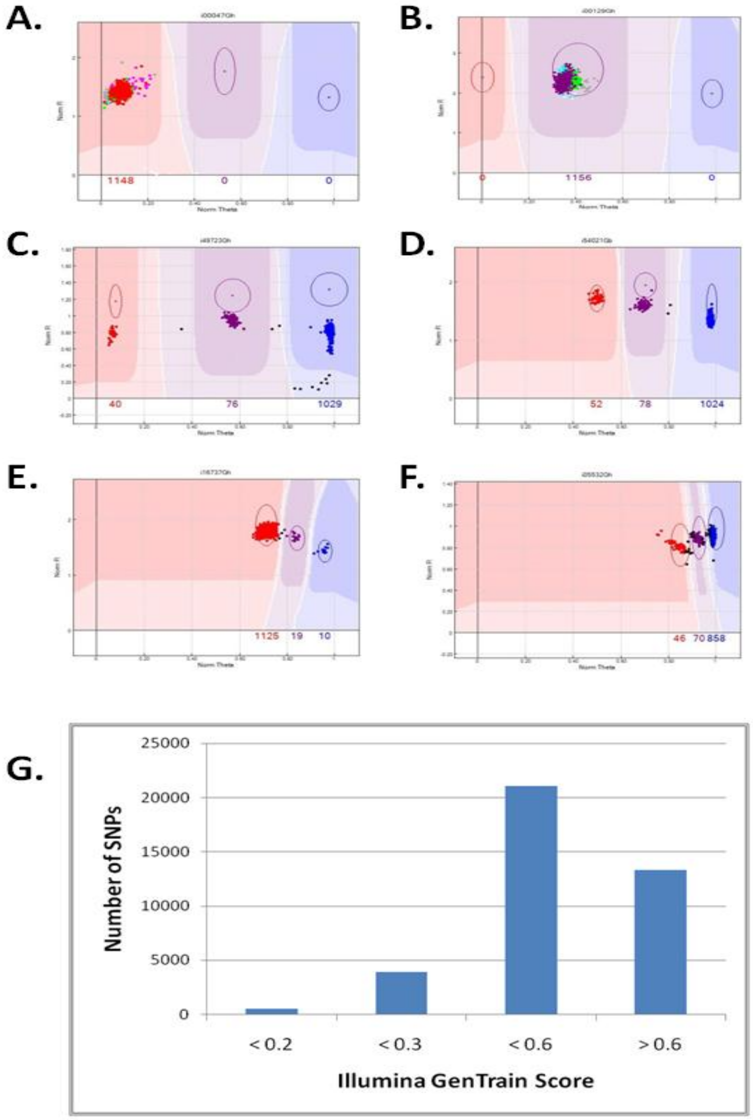


Figure 4.3 Classification of scorable SNP markers according to Illumina GenTrain score. NormTheta or relative amount of each of the two fluorophore signals is plotted on the X-axis, while NormR or signal intensity is plotted on the Y-axis. **A.**) Monomorphic marker, **B.**) Intergenomic or homeo-SNP marker, [C-F] Classification of polymorphic markers based on Illumina GenTrain score. **C.**) Genome specific marker representing a single polymorphic locus with GenTrain score >0.6, **D.**) Marker with GenTrain score 0.30-0.59 on half the plot representing two genomes, one monomorphic and one polymorphic locus, **E.**) Marker with GenTrain score 0.21-0.29 representing multiple monomorphic loci and one polymorphic locus, **F.**) Marker with GenTrain score less than 0.20 representing many monomorphic loci and one polymorphic locus, **G.**) Distribution of cluster types in polymorphic markers based on GenTrain score.

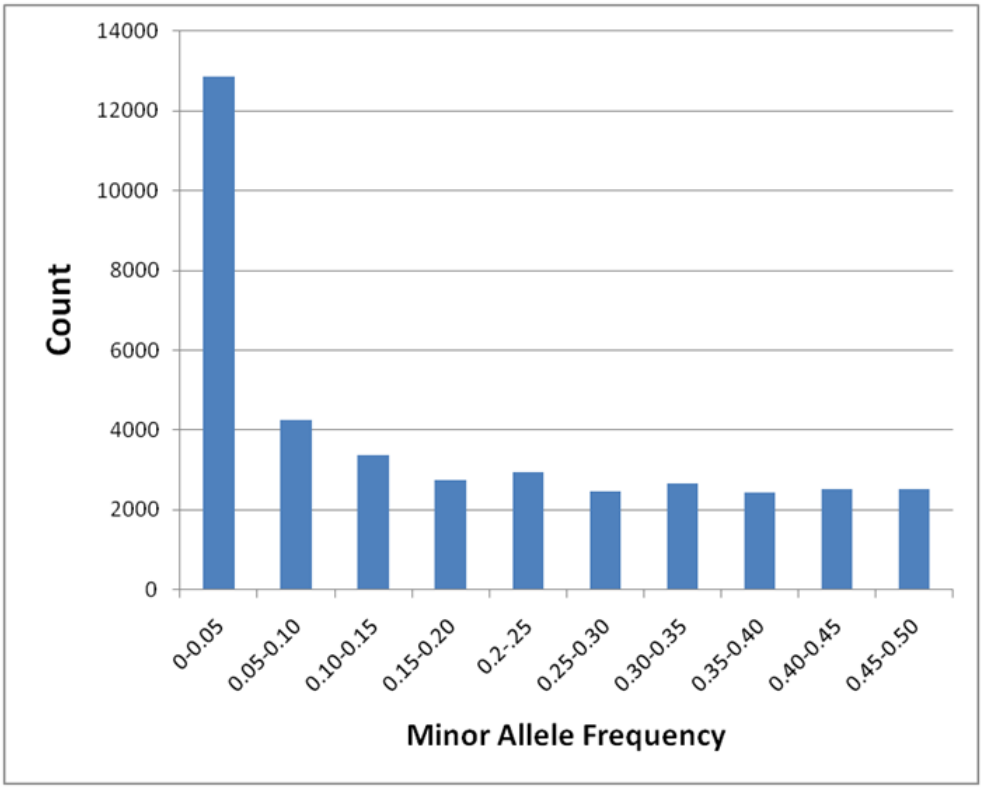


Figure 4.4 Distribution of minor allele frequencies of all polymorphic SNPs on the CottonSNP63K array. Minor allele frequencies were determined using only inbred line samples, mapping samples and other non-inbred line samples used for cluser file development were excluded from this analysis.

Table 4.4 shows the distribution of the final markers and their success rates across the designed data sets. While most data sets had a success rate of 50% or greater, some had a much lower success rate. Discovery sets that did not have previous validation or had limited validation with PCR-based assays frequently resulted in lower success rates, e.g. CSIR-NBRI and USDA-Set 1. In the CSIR-NBRI data set low success is likely due to homeo-SNPs as it showed the highest percentage of markers classified as (33.82%). The other data set that had a low success rate of 8.65% was the USDA-Set 1 containing markers previously mapped by GBS (Gore, Fang et al. 2014). While the CSIR-NBRI data set showed elevated levels of intergenomic markers, the USDA-Set 1 was primarily found to have monomorphic markers (77.88%). As these markers had been previously mapped and their map positions correlate with alignment positions on JGI D_5 genome (Gore, Fang et al. 2014), these are likely true markers that are of very low MAF or specific to the individual bi-parental population used in their identification (Gore, Percy et al. 2012). While lines for the parents of the population were included in the study, this population has been shown to contain non-parental derived alleles (Gore, Percy et al. 2012). Thus without lines that represent the non-parental derived alleles, polymorphism would not be able to be identified and this would lead to the monomorphic classification of these markers.

Table 4.4 Distribution of classified SNPs across the discovery sets and success rates of these SNPs on the CottonSNP63K array.

Data Set	SNPs on Array	Failed		Successful Assays (#)			Success Rate (%)
		#	%	Monomorphic	Intergenomic	Polymorphic	
Brigham Young University	185	23	12.43%	16	5	141	76.22%
CSIR-NBRI	343	41	11.95%	84	116	102	29.74%
USDA-Set1	104	10	9.62%	81	4	9	8.65%
UC-Davis/TAMU GH RNA-Seq	938	153	16.31%	81	66	638	68.02%
USDA-Set2	2,223	474	21.32%	193	372	1,184	53.26%
CSIRO	17,230	2,048	11.89%	4,325	772	10,085	58.53%
TAMU/UC-Davis Intra Genomic - Set 1	23,418	3,509	14.98%	4,639	3,565	11,705	49.98%
UC-Davis/TAMU Intra Genomic - Set 2	445	48	10.79%	5	41	351	78.88%
DOW AgroSciences	218	13	5.96%	1	0	204	93.58%
UC-Davis Inter	143	10	6.99%	0	1	132	92.31%
CIRAD	145	20	13.79%	14	27	84	57.93%
TAMU/UC-Davis Inter RNA-Seq	13,055	913	6.99%	374	307	11,461	87.79%
TAMU/UC-Davis Inter Genomic	4,611	595	12.90%	501	789	2,726	59.12%
Total	**63,058**	**7,857**	**12.46%**	**10,314**	**6,065**	**38,822**	**61.57%**

Reproducibility and call rate in different cotton samples

Replicates of samples were analyzed in order to determine reproducibility across genotyping runs as well as reproducibility across seed sources. To do this, four different types of replicates were used: 1) technical replicates running the same DNA on different genotyping runs, 2) individual plants from the same seed source, 3) individual plants from different seed sources, and 4) pooled DNA from multiple plants from the same seed source. The three technical replicates (three samples from each *G. hirsutum* line TM-1, *G. barbadense* line 3-79 and their F_1) showed negligible inconsistencies between runs with average similarity of 99.93% ± 0.0007. Therefore allele calls are highly reproducible across runs. Individual plants from the same seed source showed a range of similarity from 100% for Coker 315 provided by CSIRO to 73.39% for Lu Mien 14 samples provided by CSIRO. A larger inconsistency was seen with duplicated samples from different seed sources. The amount of inconsistency varied considerably across different lines, from 66.61% for VIR-6615/MCU-5 samples provided by USDA-ARS (College Station) and CSIRO to 99.36% for Stoneville 474 samples provided by USDA-ARS (New Orleans) and CIRAD. When DNA of a pool of individual plants developed from the same seed source was analyzed, the percentage of similarity varied between lines from 79.48% for LuMien 14 to 99.84% for Sicot 189. Rates determined for individual lines and replications of different types are shown in **Table 4.5**.

Table 4.5 Percent similarities for technical and biological replicates of lines.

Line	Percent Similarity				Percent Residual Heterozygosity in Pools
	Technical Replicates	Individual Plants		Line Pool	
		Same Seed Source	Diff. Seed Source		
TM-1	100.00*	-	89.77	-	-
3-79	99.90 (±0.0006)	-	-	-	-
F_1(TM-1x3-79)	99.87 (±0.0006)	-	-	-	-
Coker 315	-	100.00	89.32	99.67 (±1.83E-5)	0.33
Delta Opal	-	97.86	-	96.16 (±0.0015)	3.51
Deltapine 16	-	96.23	95.54	94.62	4.96
Deltapine 90	-	99.74	95.68	99.47 (±0.0023)	0.57
LuMien 14	-	73.39	-	79.48 (±0.1688)	19.40
MCU-5	-	99.52 (±0.0020)	66.61	99.42 (±0.0019)	0.64
Namcala	-	97.17	-	95.92	2.72
Riverina Poplar	-	-	-	80.52	10.82
Sicala 40/Fibermax 966	-	-	97.59 (±0.0202)	99.25	0.57
Sicala V-2/Fibermax 989	-	-	-	99.60	0.38
Sicot 189	-	-	-	99.84	0.15
Sicot 71	-	-	-	98.92	1.06
Sicot 81	-	-	-	99.29	0.67
Sicot F-1	-	-	-	92.53	6.63
F_1(TM-1xim)	-	99.97	-	-	-
F_1(DP5690xLi2)	-	99.87	-	-	-
F_1(STV474xHS26)	-	94.58	-	-	-
Acala Maxxa	-	-	88.16	-	-
Stoneville 474	-	-	99.36	-	-
Guazuncho II	-	-	89.19	-	-
Coker 312	-	-	84.20	-	-
Tamcot SP37	-	-	89.28	-	-

Standard deviations are listed for comparisons with three samples, where no standard deviation is listed comparisons are between two samples.
*Complete identity, i.e. no standard deviation between the three samples.

Genetic map construction

An intra-specific map was generated from 93 F_2 samples from a cross between lines Phytogen 72 and Stoneville 474. A total of 7,171 SNP markers were mapped in 26 linkage groups representing 3,499 cM (**Figure 4.5**). These represent 6,938 markers from the *G. hirsutum* set and 235 markers from the other species sets. An average of 254 markers per linkage group were mapped on A-subgenome chromosomes and 298 markers per linkage group of the D-subgenome. These correspond to an overall average of 55 bins with 4.98 markers/bin per chromosome, 51 bins with 4.59 markers/bin for A-subgenome chromosomes and 60 bins with 5.40 markers/bin for D-subgenome chromosomes.

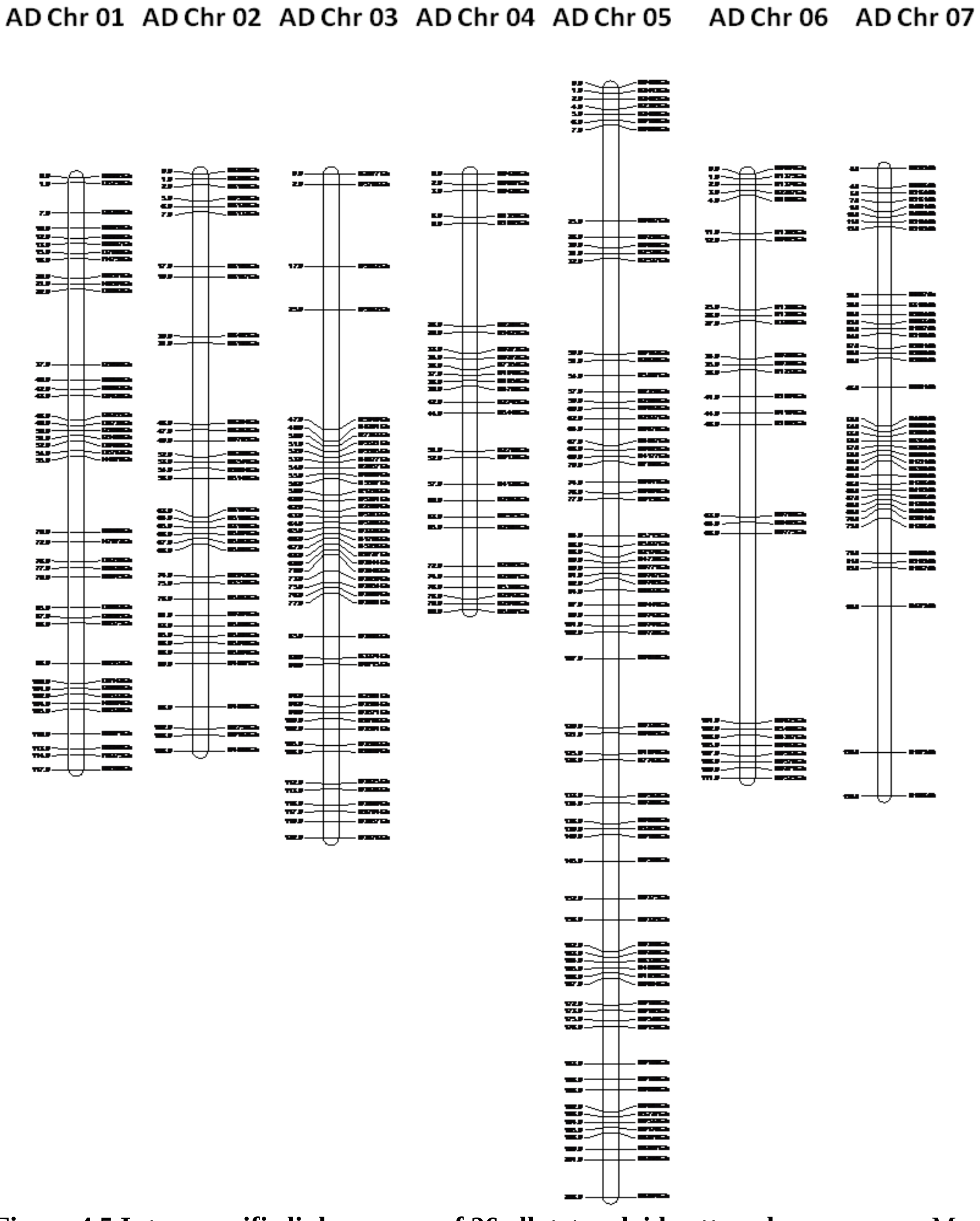


Figure 4.5 Intra-specific linkage map of 26 allotetraploid cotton chromosomes. Map determined using 93 F2 individuals from Phytogen 72 by Stoneville 474 parents. Only one marker is listed on the right per Kosambi centiMorgan (cM) on the left, even if there were more markers co-segregating. Chromosomes are listed based on AD chromosome number.

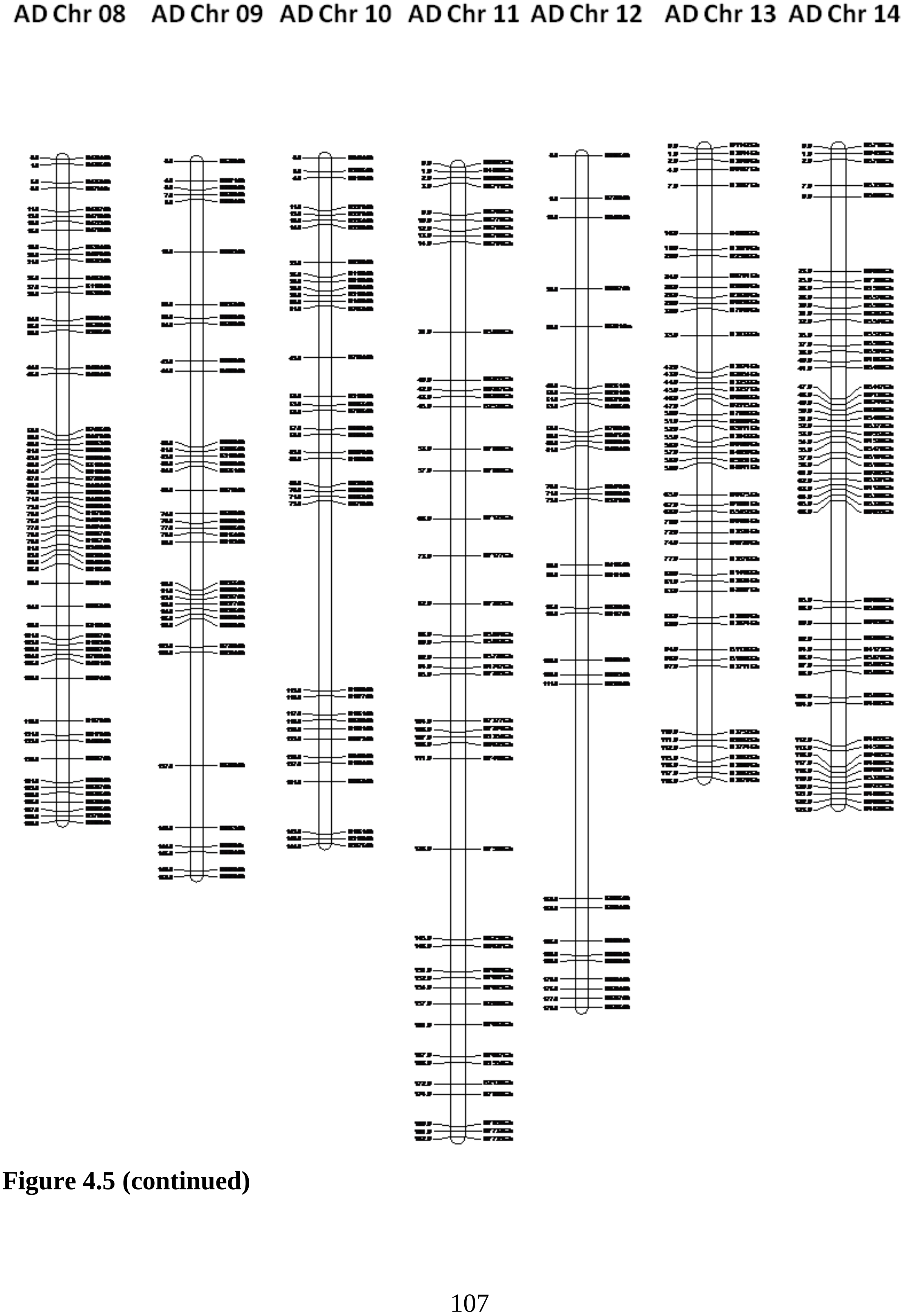


Figure 4.5 (continued)

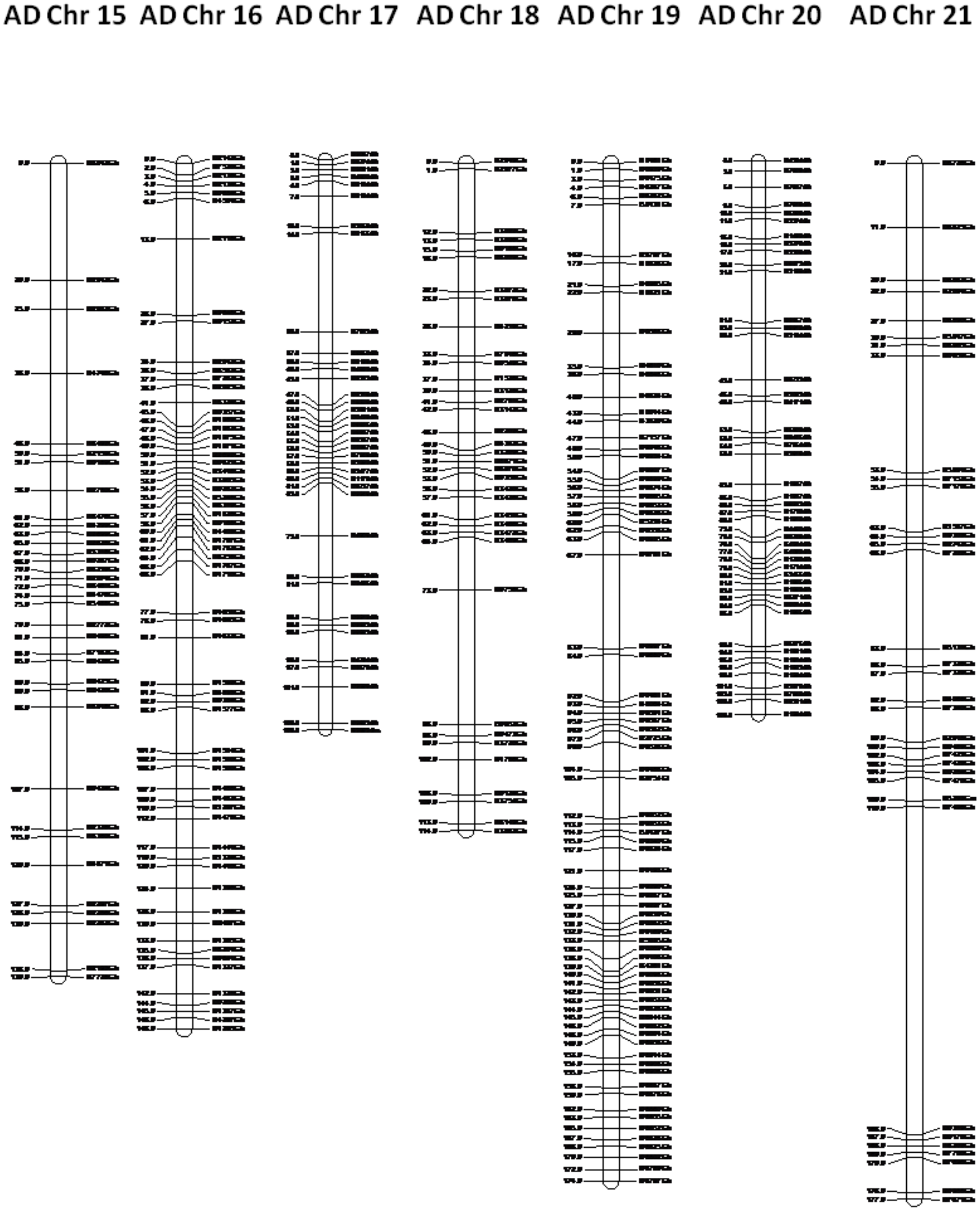


Figure 4.5 (continued)

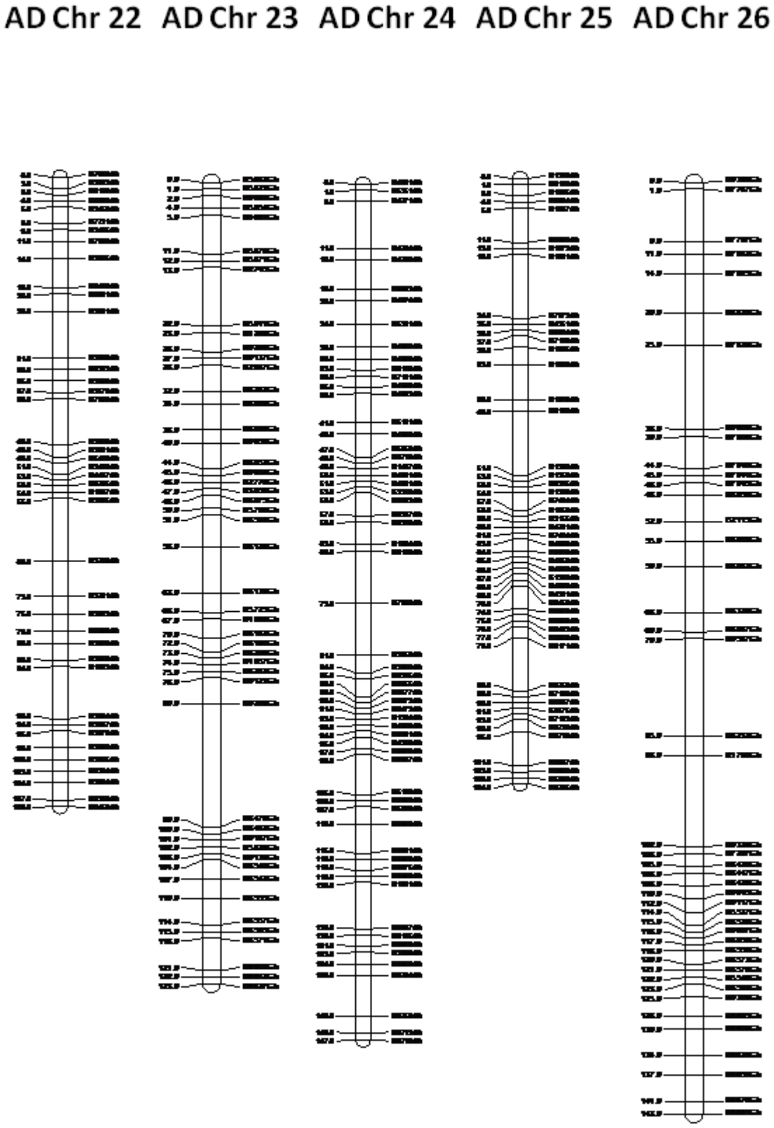


Figure 4.5 (continued)

An initial inter-specific map was generated from 118 F_2 samples from a single inter-specific cross between *G. hirsutum* standard line TM-1 and *G. barbadense* line 3-79. A total of 19,198 markers mapped into 26 linkage groups initially representing 4,439.6 cM. Markers were largely distributed across the genome with a few moderately sized gaps. Correlation analysis between the initial inter-specific map and the intra-specific map showed that the moderately sized gaps were associated with inverted marker orders compared to the intra-specific map in nine linkage groups corresponding to chromosomes (Chr06, Chr13, Chr15, Chr18, Chr20, Chr21, Chr22, Chr23, and Chr26). Linkage disequilibrium and recombination plots of these linkage groups showed incorrect ordering near the gaps (See Chr06 example in **Figure 4.6**). The most likely correct orientation is found in the intra-specific linkage groups, where the gaps are smaller. For example in **Figure 4.6**, it is likely that the inter-specific map is inverted due to low linkage across the gap while the intra-specific map does not have a corresponding gap and is likely the correct orientation. To correct this issue, the intra-specific map was used as a framework map to reorder the problematic inter-specific linkage groups.

The final map includes the reordered linkage groups and contains 19,191 markers mapped to 26 linkage groups that collectively encompass 3,955.1 cM (**Figure 4.7**). This map represents a reduction in size of 484.5 cM (11%) from the initial inter-specific map. The mapped SNPs represent 11,452 markers from the *G. hirsutum* set and 7,746 markers from the other species data sets (6,785 - *G. barbadense*, 614 - *G. tomentosum*, 280 - *G. mustelinum*, 39 - *G. armourianum*, and 28 - *G. longicalyx*). The map contained an average of 162 recombination bins with 4.55 markers/bin per chromosome across the linkage groups, with an average of 166 bins with 4.52 markers per bin and 159 bins with 4.60 markers per bin in the A- and D-subgenome chromosomes, respectively. The average overall number of markers per linkage group was 738, with an average of 718 per A-subgenome chromosomes and 759 per D-subgenome chromosomes.

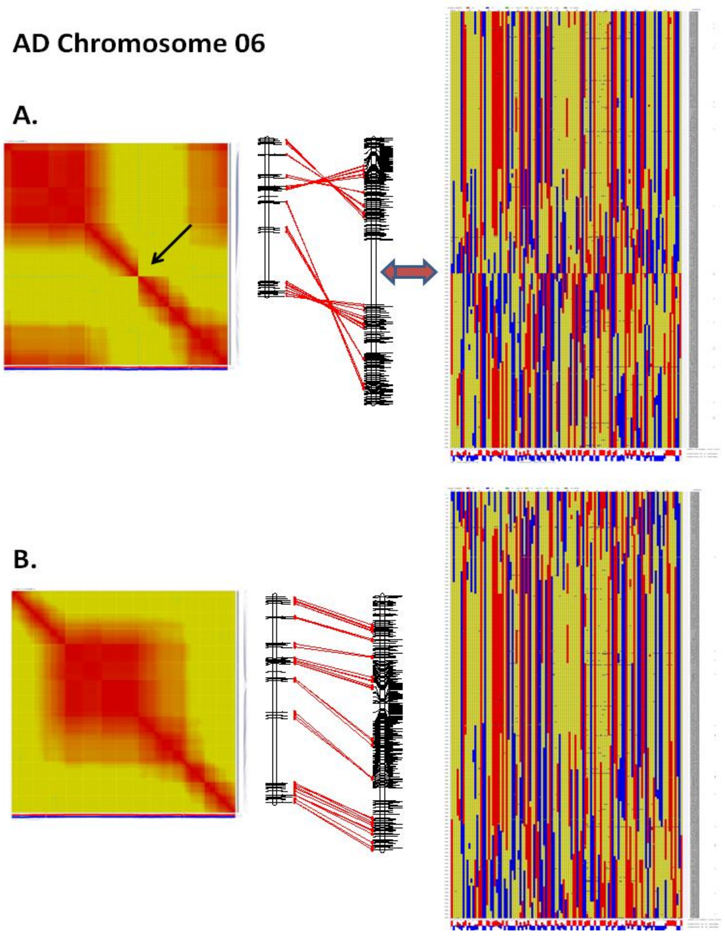


Figure 4.6 Inconsistencies between initial *de novo* inter-specific map and the intra-specific map. A.) Initial plots of inter-specific map order and correlation with intra-specific map show area of incorrect placement in center of the linkage group. B.) Corrected inter-specific linkage group and final plots.

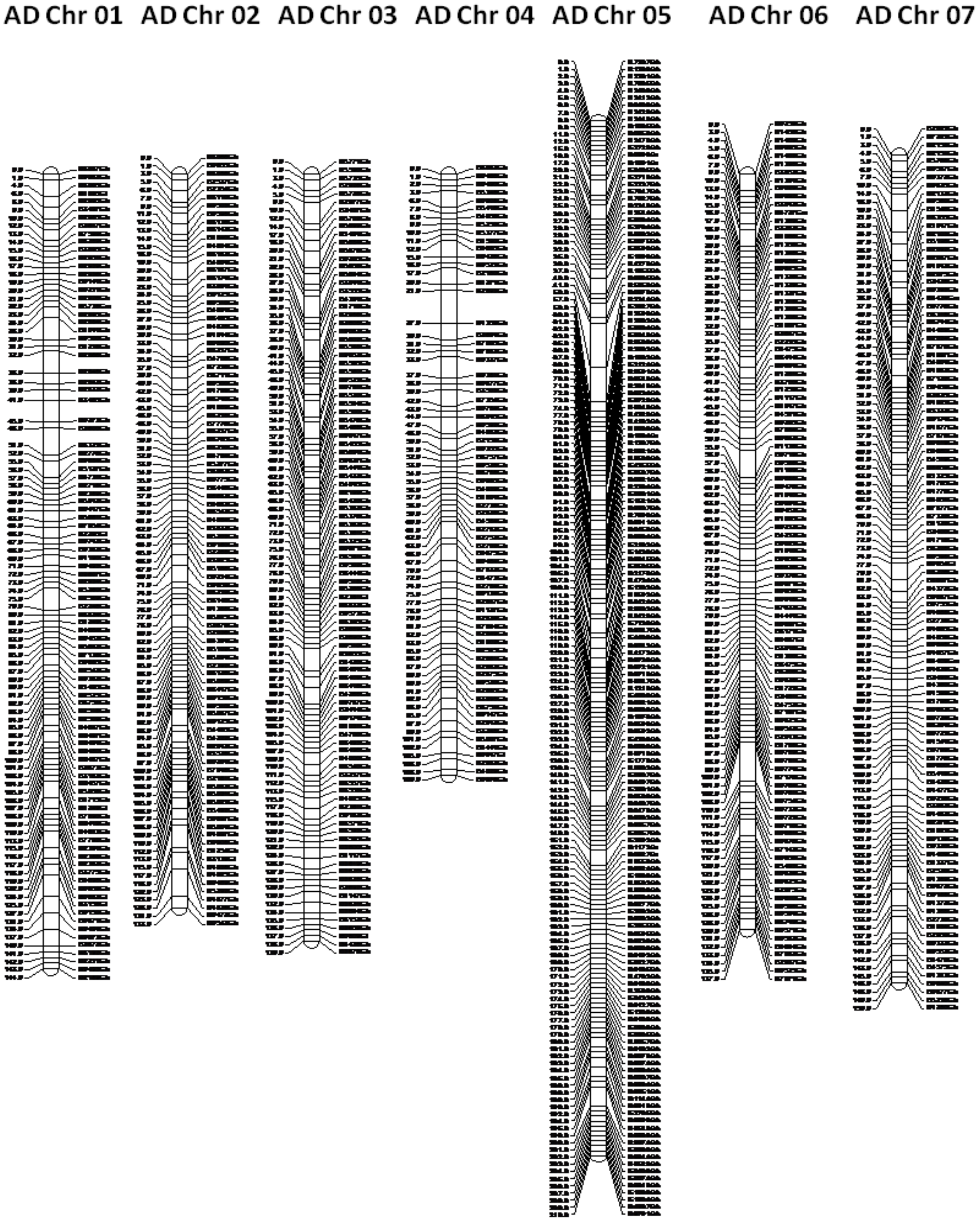


Figure 4.7 Inter-specific linkage map of 26 allotetraploid cotton chromosomes. Map determined using 118 F2 individuals from *G. hirsutum* genetic standard line Texas Marker -1 by *G. barbadense* line 3-79 parents. One marker listed on the right per Kosambi centiMorgan (cM) on the left as in **Figure 4.6**. Chromosomes are listed based on AD chromosome number.

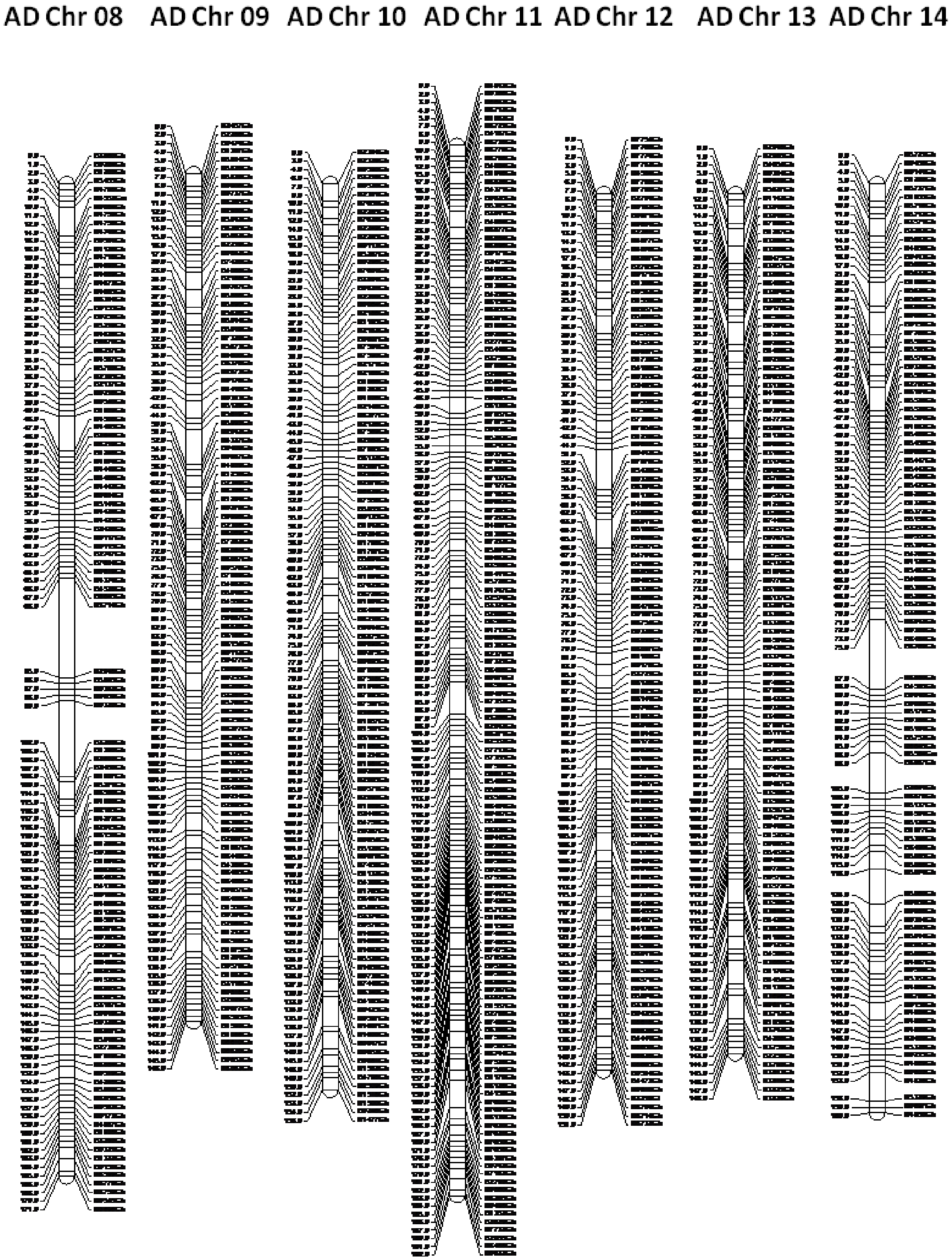


Figure 4.7 (continued)

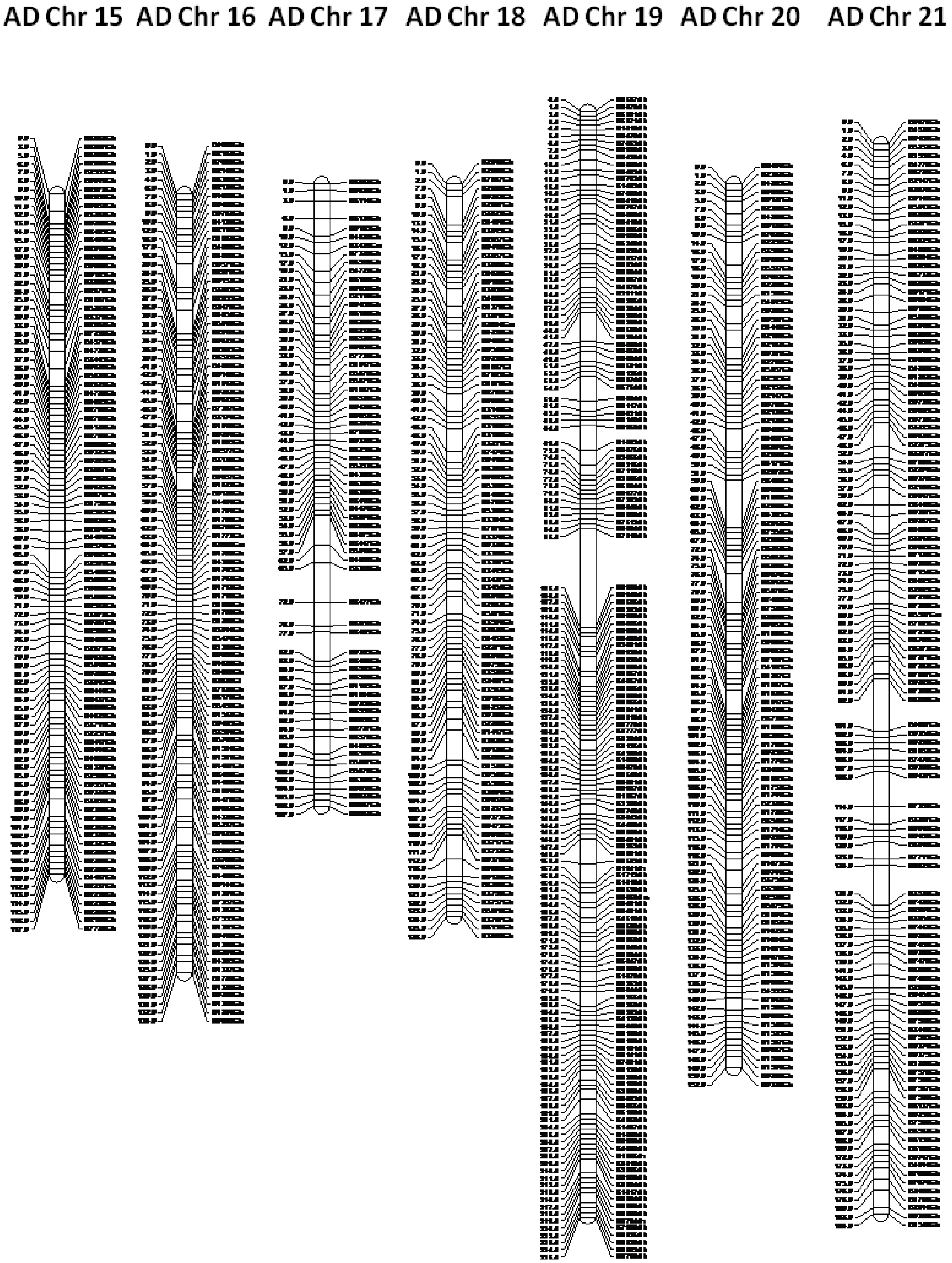


Figure 4.7 (continued)

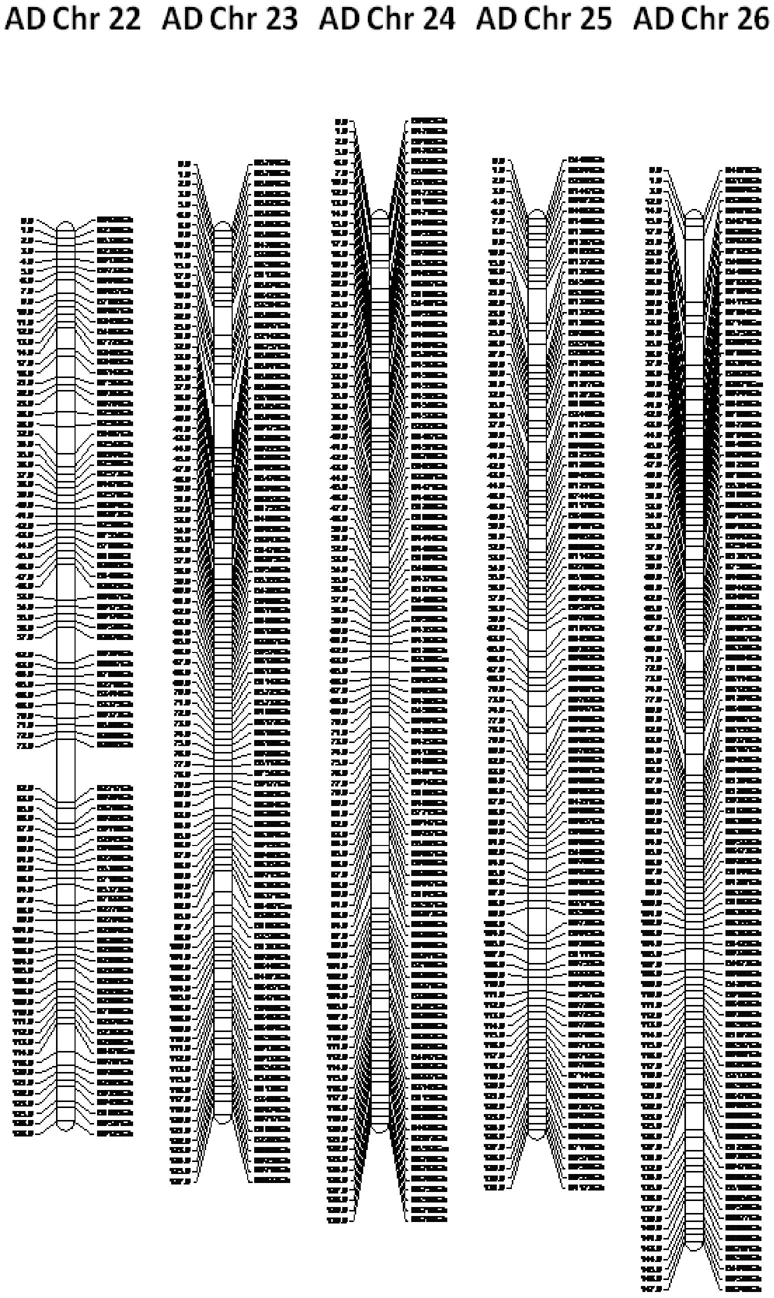


Figure 4.7 (continued)

In the intra-specific and inter-specific maps, the observed map lengths (cM) for the A-subgenome chromosomes (137.5cM, 152.8cM) were on average very similar to but slightly larger than for the D-subgenome chromosomes (131.7 cM, 143.7 cM), respectively. For cotton chromosomes, accurate estimates and comparisons of genetic size with physical distance, e.g., cM/Mbp, are not yet available. Accurate cytological estimates for individual allotetraploid chromosome sizes are not available. While a BAC-by-BAC reference sequence for the D_5 is available, highly repetitive regions are under-represented, including subtelomeric and large peri-centromeric regions. So, a D_5-based estimate would be limited in scope and would underestimate physical size. However, the genomes of the closest related extant diploids, G. arboreum and G. raimondii, have been estimated to be 1600 and 800 Mb, respectively; sizes of those genomes and chromosomes correspond well cytologically with absolute and relative sizes of chromosomes in the G. hirsutum subgenomes. Using the G. arboreum and G. raimondii estimates as a guide for A-subgenome and D-subgenome sizes, one centimorgan in the intra-specific map represents an average of 895 Kb in A-subgenome chromosomes and 467 Kb in D-subgenome chromosomes. In the inter-specific map, a centimorgan represents an average physical distance of 805 Kb in A-subgenome chromosomes and 428 Kb in D-subgenome chromosomes. While sizes and physical distances in A- and D-subgenome chromosomes differ about 2-fold, the rate of crossing over does not, reflecting that most crossovers occur in non-heterochromatic regions or gene-space.

Segregation distortion has been previous reported in cotton populations and not unexpectedly a number of markers in both the intra-specific and inter-specific maps showed significant distortion from expected F_2 segregation ratios. In the intra-specific map, 612 markers of the 7,171 (8.5%) mapped markers showed significant segregation distortion ($P<0.05$). Of these distorted markers 256 (41.8%) show a decrease in amount of heterozygotes. Overall when one parental allele is favored, generally an excess of the female parental allele (Phytogen 72) was observed at a 3.45 to 1 ratio. In the inter-specific map, 3,475 markers of the 19,191 (18.1%) mapped markers showed significant segregation distortion ($P<0.05$). Unlike the intra-specific markers, when one parental allele is favored, there is not a large bias towards one parent and a favored ratio of 1.13 is observed for male parent plant (*G. barbadense*) to female parent (*G. hirsutum*). A skew towards the *G. barbadense* parent was also observed in Lacape et al. (2003). The amounts of distorted markers in these populations are within the range of previous reported studies from 8.5% to greater than 50% (Lacape, Nguyen et al. 2003; Shen, Guo et al. 2007).

Distribution of recombination events

The high density of markers mapped in this study allowed the identification of almost all of the recombination events present in each F_2 individual. We calculated the number of crossovers per F_2 using all markers in both the intra-specific and inter-specific linkage maps. In the intra-specific map the number of crossovers across all chromosomes of an individual varied from 44 to 97 with an average of 67.5, ignoring two individuals with an abnormally high number of crossovers on a couple of linkage groups. These individuals were ignored as while they showed a normal number of crossovers across most chromosomes they had much higher >2X the number of crossovers as all other samples for one or more linkage groups, one individual had 34 crossovers on Chr13 and the other individual had 24, 34 and 38 crossovers on Chr05, Chr19, and Chr14, respectively. An average of 2.65 crossovers per linkage group occurred across all individuals. Distribution of crossovers in the intra-specific population across linkage groups is shown in **Figure 4.8a**.

In the inter-specific map, the number of crossovers per individual ranged from 54 to 97 with an average of 75.8 (ignoring a single individual that had an abnormally high number of crossovers in AD14). Similar to the intra-specific mapping population, there was an average of 2.92 crossovers per linkage group across individuals. The distribution of crossovers in the inter-specific population across linkage groups is shown in **Figure 4.8b**.

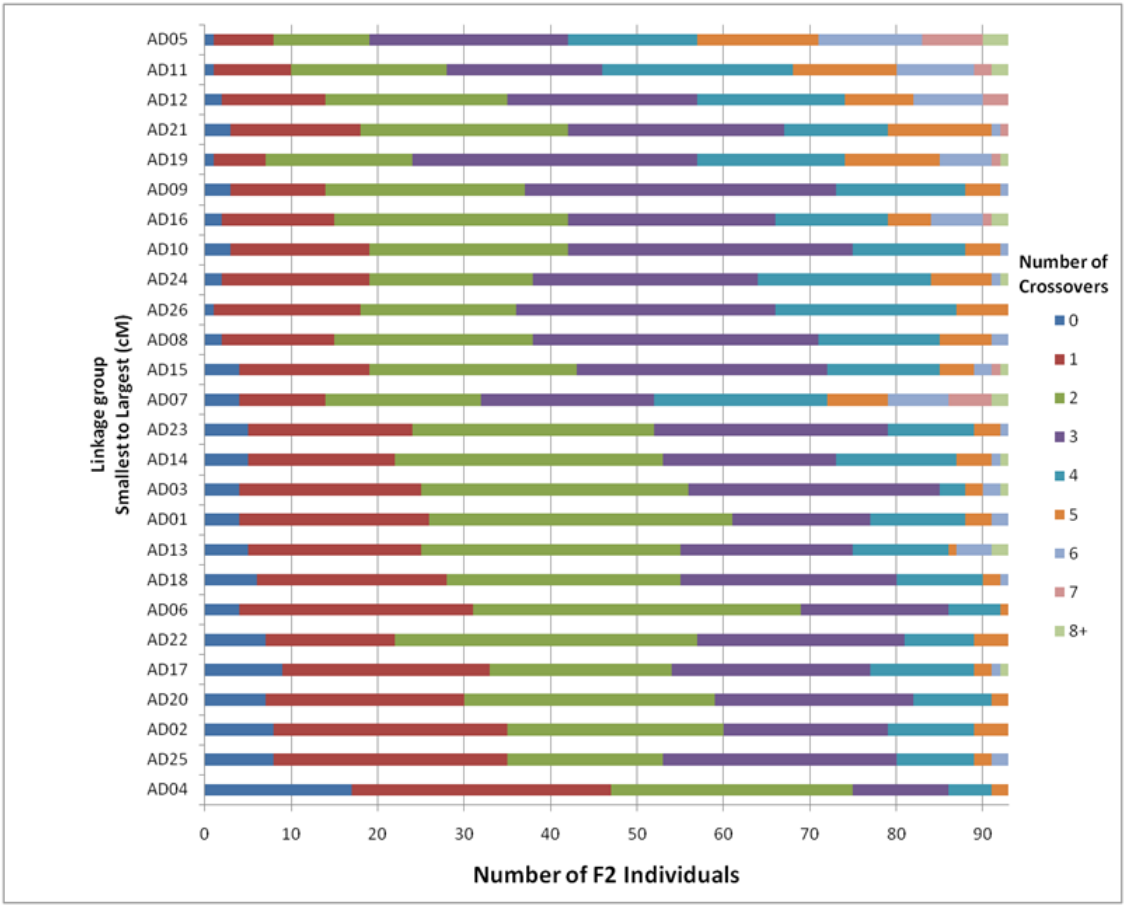


Figure 4.8 Frequency distribution of the number of crossovers. Numbers of crossovers detected for each F2 individual per chromosome (0 to >8) are displayed chromatically for each linkage group, which are organized by genetic size (longest at top, shortest at bottom). **A.**) Distribution of crossovers in the intra-specific mapping population. **B.**) Distribution of crossovers in the inter-specific mapping population.

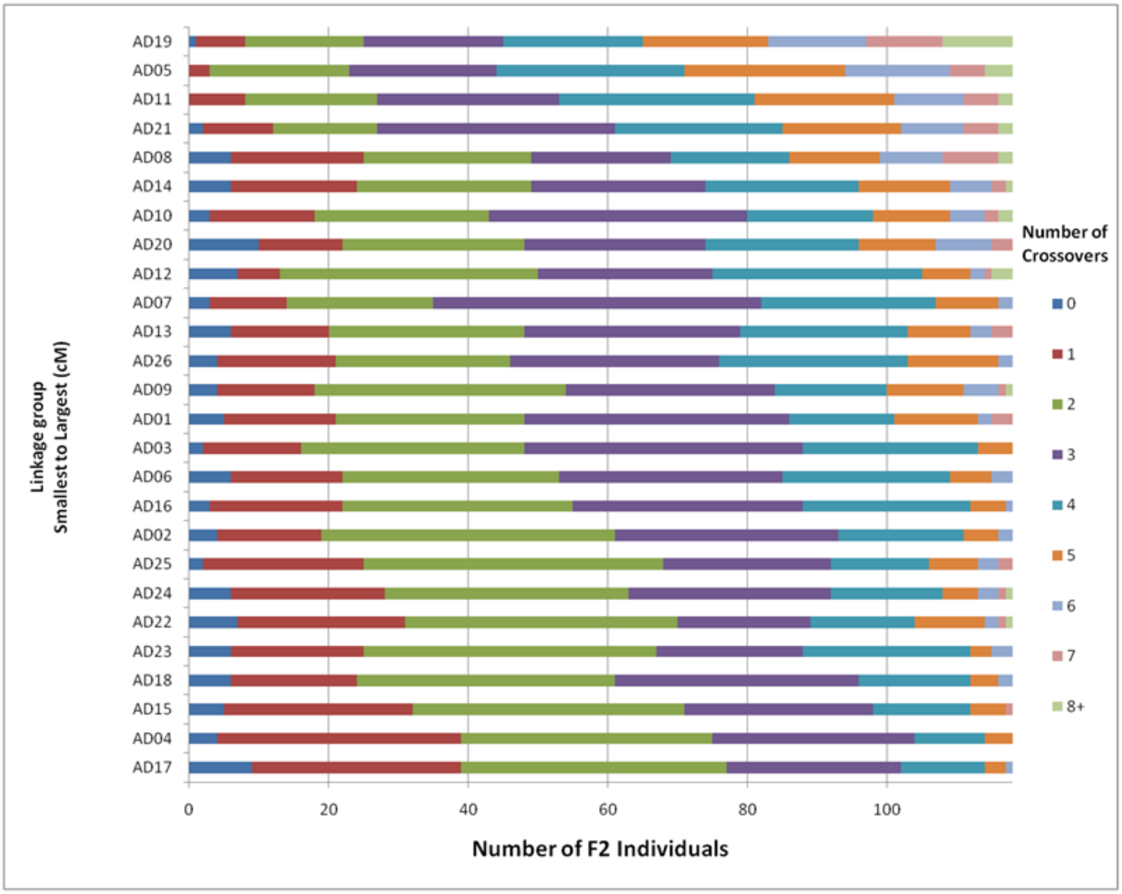


Figure 4.8 (continued)

Analyses of synteny

All of the 70,000 putative SNP markers that were used for production of the CottonSNP63K array were aligned to the D_5 reference genome using BWA in Galaxy; 59.7% of the markers produced alignments to the reference. Synteny was detected for a total of 4,521 and 12,027 mapped SNPs for the intra-specific and inter-specific maps, respectively. Dot plots of the linkage maps versus the D_5 reference genome show high collinearity across linkage groups and chromosomes as expected, except for four A-subgenome chromosomes, which are known to contain translocations relative to D-chromosomes. Translocations involving chromosomes 2, 3, 4, and 5 were identified in both maps and are indicated in **Figures 4.9a** and **4.9b**.

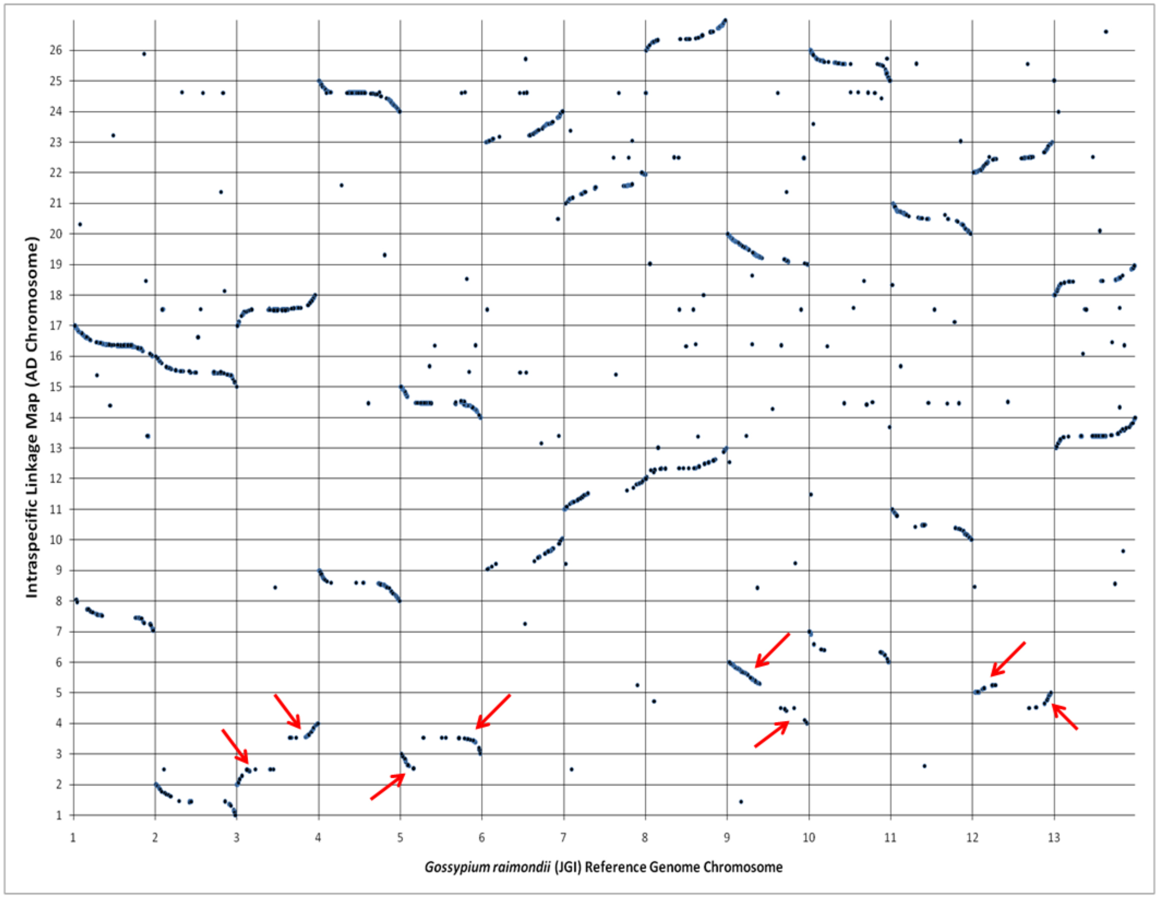


Figure 4.9 Dot plot of the syntenic positions of SNP markers in the allotetraploid linkage maps versus the JGI *G. raimondii* reference genome. The 26 allotetraploid chromosomes are shown on the y-axis and the 13 chromosomes of *G. raimondii* are shown on the x-axis. Red arrows indicate translocation events relative to *G. raimondii*. **A.**) Intra-specific linkage map displaying positions of 4,521 mapped SNP in *G. hirsutum* with alignments to *G. raimondii*. **B.**) Inter-specific linkage map (*G. hirsutum* genetic standard line Texas Marker -1 by *G. barbadense* line 3-79) displaying positions of 12,027 mapped SNP with alignments to *G. raimondii*.

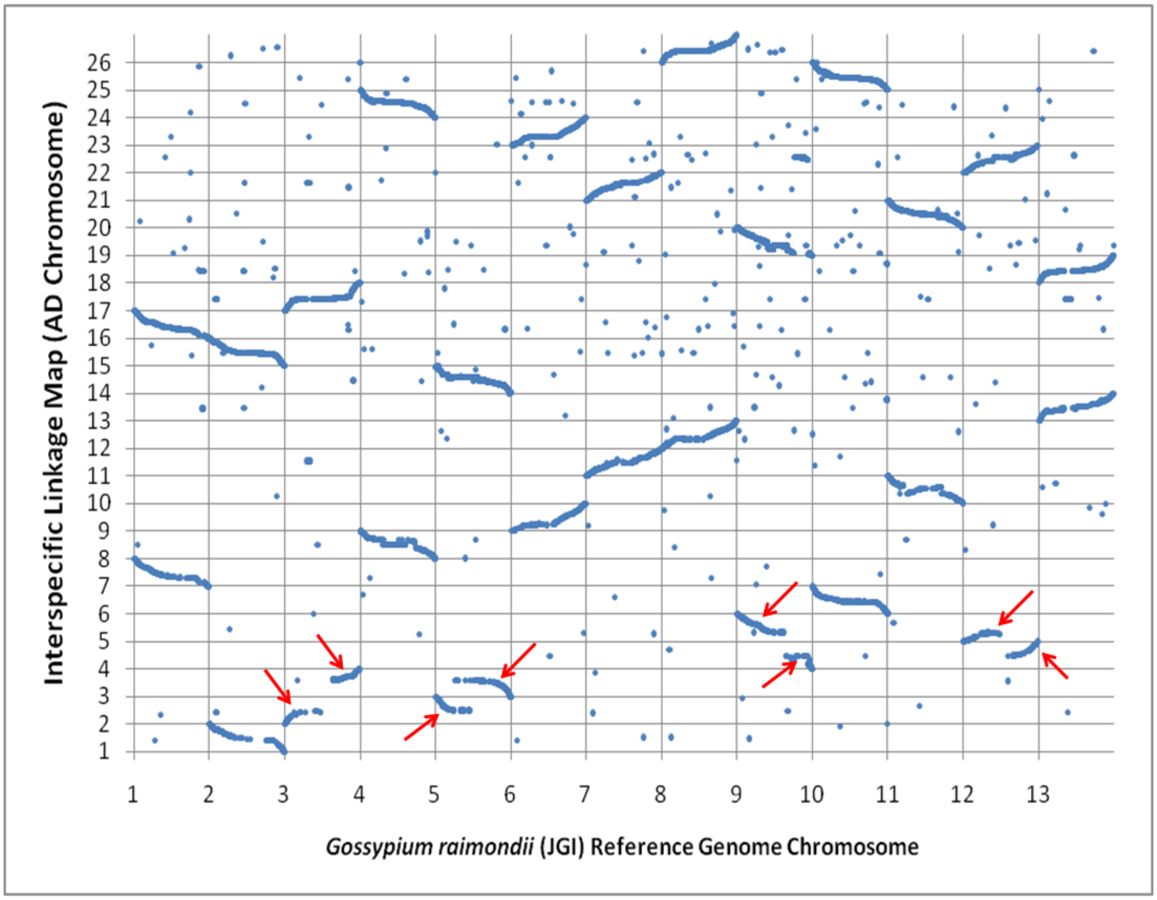


Figure 4.9 (continued)

When all markers were aligned to the A_2 draft genome, 61.8% of the markers aligned to the reference. A total of 3,863 and 11,344 SNPs mapped that were included in the intra-specific and inter-specific maps, respectively. While a larger percentage of the SNPs mapped to A_2 compared to D_5, expected collinearity was not observed when these A_2 alignment positions were plotted against position in the linkage maps (**Figures 4.10a and 4.10b**) like was seen with the high-quality BAC-by-BAC derived D_5 reference genome. The strong collinearity with the high-quality D_5 genome suggests that our genetic maps are correctly ordered while the low collinearity with the A_2 genome sequence suggests that the genome sequence may be improved using the maps produced in this effort and other genetic maps.

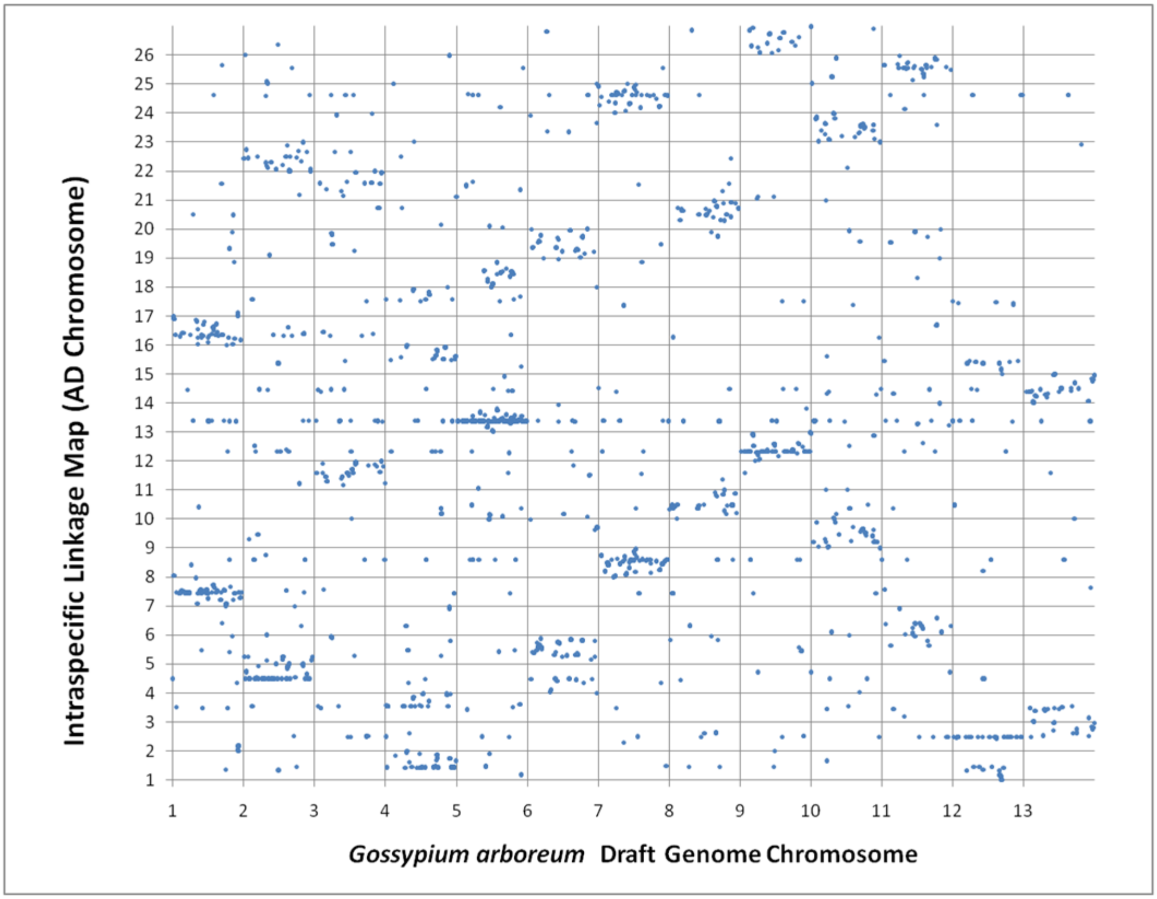


Figure 4.10 Dot plot of the syntenic positions of SNP markers in the allotetraploid linkage maps versus the BGI *G. arboreum* draft genome. The 26 allotetraploid chromosomes are shown on the y-axis and the 13 chromosomes of *G. arboreum* are shown on the x-axis. **A.**) Intra-specific linkage map displaying positions of 3,863 mapped SNP in *G. hirsutum* with alignments to *G. arboreum*. **B.**) Inter-specific linkage map (*G. hirsutum* genetic standard line Texas Marker -1 by *G. barbadense* line 3-79) displaying positions of 11,344 mapped SNP with alignments to *G. arboreum*.

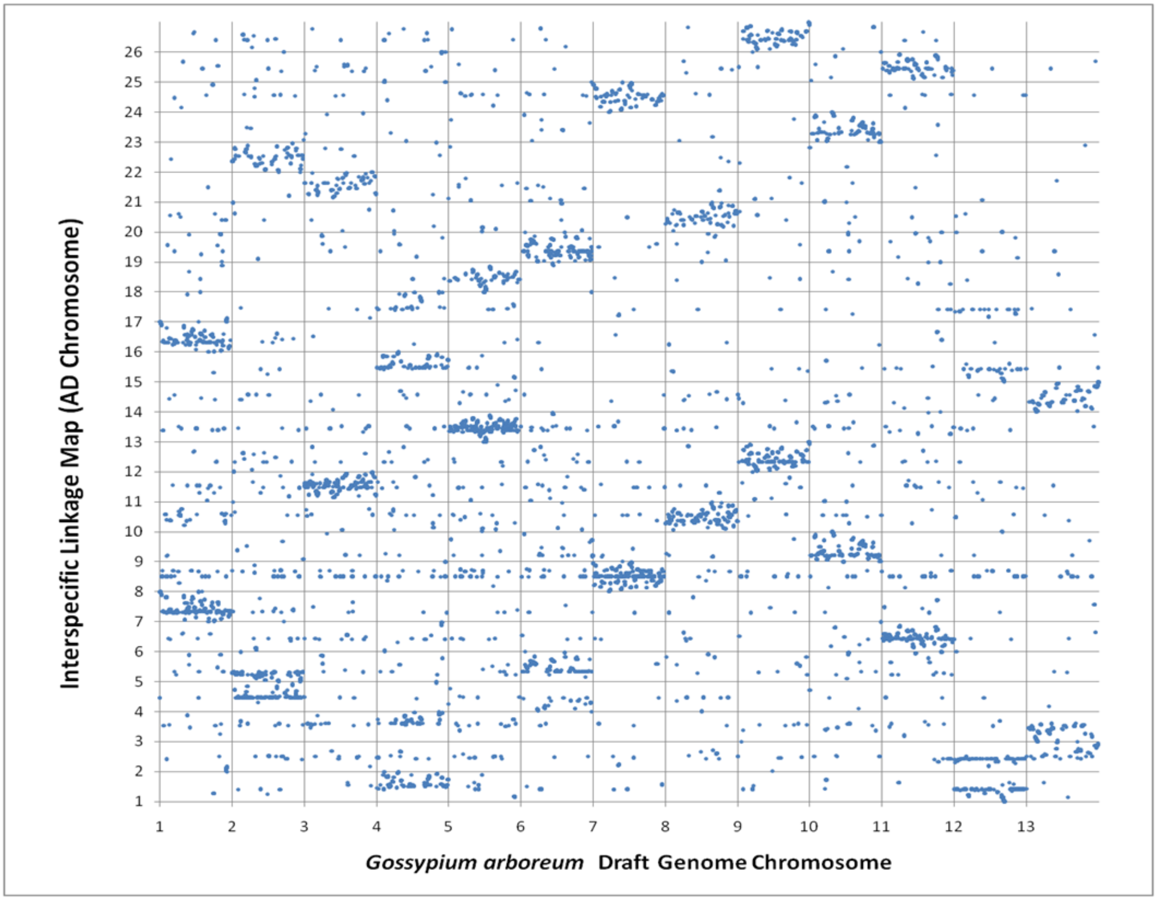


Figure 4.10 (continued)

Discussion

The inter- and intra-specific genetic maps developed with the CottonSNP63K represent the most saturated maps developed for cotton to-date. The produced intra-specific genetic map is the first saturated map for a cross between two *G. hirsutum* lines. While one other effort has come close to generating a number of linkage groups equal to the 26 cotton chromosomes (Zhang, Hu et al. 2009), to the best of our knowledge the intra-specific map developed here is the first that associates into 26 linkage groups corresponding to the number of cotton chromosomes. The array and maps provide a foundation for fine mapping and genetic dissection of agronomically and economically important traits, and facilitate the identification of causative genes underlying quantitative trait loci. This new tool will also foster map-based cloning and genome assembly efforts, as well as contribute to advancements in marker-assisted selection and genomic selection in breeding programs. While our initial inter-specific map was very close to the estimated tetraploid map size of ~4,500 cM suggested by Rong et al. (2004), it showed sizeable gaps and discordant LD patterns. Once corrected, the size of both maps was well below the estimated value of ~4,500 cM at 3,955.1 cM for the inter-specific map and 3,499cM for the intra-specific map. A decrease in map size is expected when higher-density maps are generated, due to the improved ability to distinguish scoring errors from double recombinant events with high density markers (Rong, Abbey et al. 2004). While the sizes of the new inter- and intra-specific maps are lower than expected, they both fall well within the range of previous map sizes of published inter-specific maps 3,380 – 5,115 cM (Blenda, Fang et al. 2012; Yu, Kohel et al. 2012; Shi, Li et al. 2014) and within the 2,061 to 4,448 cM of reported intra-specific maps (Rong, Abbey et al. 2004; Zhang, Zhang et al. 2012; Gore, Fang et al. 2014; Tang, Teng et al. 2014).

The mapping of a large number of markers from different discovery sets utilizing a large variety of different cotton lines, primarily cultivars as shown in **Table 4.2**, along with a moderate percentage of markers showing intermediate to high MAFs suggests high transferability of markers across different sets of cotton germplasm. In order to have a wide applicability across cotton samples, the array was designed to provide the most comprehensive tool to-date for cotton researchers and included SNPs that were discovered from technologies targeting both gene regions and genomic regions. This approach was applied so that the inclusion of markers outside of genes would generate a more even distribution across chromosomes since genic markers are unevenly distributed across chromosomes (Hulse-Kemp, Ashrafi et al. 2014). As the majority of the gene-based markers were provided by the CSIRO discovery set using the D_5 sequence assembly as a reference (Zhu, Spriggs et al. 2014), many A-subgenome-specific genes may be under represented in the gene-based marker set. Fortunately, genomic markers located in LD with a gene of interest can also be of value in genome-wide association studies, and thus help compensate for the subgenomic bias among genic intra-specific markers. As the TAMU/UC-Davis Intra Genomic – Set 1 and TAMU/UC-Davis Inter Genomic sets are markers derived from BAC-end sequences (Hulse-Kemp et al. – submitted), these markers will provide an avenue for direct map-based cloning and fine-mapping of identified regions of interest through direct integration of the array with BAC-based resources.

The array provides cotton research groups with a standardized set of markers that can be used to localize and replicate important previously identified markers, and it will also allow for investigating different types of sample inconsistencies. The replication analysis performed here showed that while technical DNA replicates show very high consistency, different levels of variability were seen with other replication types. As found in maize (Yan, Shah et al. 2009) the array analysis revealed considerable variation within some cotton lines not only between different seed sources, but also within the same seed source. Notably two different TM-1 samples, typically used as a *G. hirsutum* genetic standard line (Kohel, Richmond et al. 1970), show only 89.77 percent similarity. Hinze et al. (2015) found similar inconsistencies using SSRs in the US National Cotton Germplasm Collection where accessions with a common name may or may not have been identical while other accessions with no reason for similarity were genetically identical. Such inconsistencies can create difficulties when trying to interpret or compare studies utilizing samples bearing a common name. The CottonSNP63K provides an efficient way to characterize these inconsistencies as it has negligible inconsistencies with reproducibility, thus the similarity differences discovered between seed sources is due to variation of the sources not a problem with genotyping on the array. The amount of residual heterozygosity within the same seed source shows that some lines are more heterogeneous than others, and comparisons across studies using heterogeneous lines will be less reproducible. On the other hand, residual heterozygosity in lines also will offer direct avenues for fine mapping of genes in primarily isogenic backgrounds (Truco, Ashrafi et al. 2013) as well as provide additional diversity within *G. hirsutum* that has been noted to have low levels of diversity (Tyagi, Gore et al. 2014).

Due to the low diversity among cotton lines, multiple mapping populations will be required to map all of the polymorphic SNPs on the CottonSNP63K array. The two mapping populations here were able to map a total of 22,829 total markers with 3,533 of the total SNPs which were able to be mapped in both linkage maps. Utilizing mapped loci will be very straightforward, but utilizing the other markers will be more difficult, at least until they are mapped. Syntenic analyses with available high-quality related genome sequences such as the JGI D_5 genome provide some insights into localization information in tetraploid cotton (Paterson, Wendel et al. 2012). Allotetraploid linkage groups determined here showed a linear alignment with the D_5 genome (**Figure 4.9a & 4.9b**) and were also able to identify historic reciprocal translocation events between allotetraploid chromosomes 2/3 and 4/5 (Desai, Chee et al. 2006). The breakpoints for the translocations can be roughly mapped to homeologous regions with the linkage maps to between 21.7-29.5 Mb in D_5 chromosome 3 and between 16.7-21 Mb in D_5 chromosome 5 for allotetraploid chromosomes 2/3; 40.1-46.1Mb in D_5 chromosome 9 and between 11.5-23.2Mb in D_5 chromosome 12 for allotetraploid chromosomes 4/5. It has been suggested that these translocations involved complete arms (Blenda, Fang et al. 2012) which is not ostensibly discordant to the approximate breakpoints found here. However, additional mapping would be required to pinpoint exact locations relative to the D_5 chromosomes as map information near the breakpoints is still ambiguous and overlapping at points.

Like SNP arrays developed for other crops, we found a relatively large number of markers that were difficult to score and either required manual adjustment of markers or eliminating the marker from the analysis completely. Compared to diploids the success rates for SNPs in arrays are typically lower for polyploids, due to the presence of duplicated loci in homeologous, paralagous regions, the low levels of divergence particularly between gene copies in different subgenomes, and also between paralagous regions in the same subgenomes. The 61.6% success rate found for the CottonSNP63K is quite comparable to the success rates of arrays generated for other polyploid crops, such as 61% for oat (Tinker, Chao et al. 2014) and 63% for wheat (Wang, Wong et al. 2014). Most of the markers mapped in this study were localized to a single map position, as was also the case for the majority of markers mapped with the wheat 90K array (Wang, Wong et al. 2014), indicating that most SNP assays are specific to a single locus and/or assaying segregation of a single locus. Similarly, we also found that success rates among discovery sets included on the array were highly variable (**Table 4.4**). The differences in success rates for discovery efforts as determined by this effort can provide insights for future SNP development efforts by examining parameters and pipelines used and the outcome of those discovery sets. The information provided by the array will enable future SNP discovery efforts in cotton to focus on increasing the success rate of SNP predictions.

CHAPTER V

SEQUENCE LOCALIZATION AND ENHANCEMENT OF THE ALLOTETRAPLOID COTTON PHYSICAL MAP USING WHOLE GENOME RESEQUENCING OF INTER-SPECIFIC F1 HYPO-AENUPLOID LINES

Introduction

The most important natural textile fiber crop, cotton (*Gossypium hirsutum* L.), has a polyploid genome consisting of 26 chromosome pairs (2n=4x=52) corresponding to the At and Dt subgenomes. The subgenomes have been derived from A- and D-genome ancestors roughly similar to extant diploid (2n=2x=26) species *G. arboreum* and *G. raimondii*, respectively (Wendel and Cronn 2003). Due to the redundancy in the cultivated cotton genome from the polyploidization event approximately 1-2 million years ago between related ancestors (Wendel, Brubaker et al. 2009), it has been possible to generate stable aneuploid lines lacking either a whole chromosome (monosomic) or whole chromosome arms (monotelodisomic) (Saha, Stelly et al. 2012). Isogenic versions of these aneuploid lines have subsequently been utilized as recurrent backcross parents to produce a number of whole-chromosome substitution lines in each of which a chromosome pair has been replaced by a homologous pair from a different allotetraploid *Gossypium* species. These lines have been developed using three allotetraploid species as donors, namely *G. barbadense, G. tomentosum*, and *G. mustelinum* (Stelly, Saha et al. 2005; Saha, Raska et al. 2006). Chromosome substitution (CS) lines have been used for association of important agronomic traits to specific chromosomes (Saha, Wu et al.

2004; Saha, Wu et al. 2013). CS lines have also been used in small-scale development of a few hundred SNP markers using reduced representation library sequencing (Chen, Yao et al. 2014). Aneuploid lines have also been used to generate inter-specific F_1 hypo-aneuploids by crossing monosomic *G. hirsutum* plants as seed parent with different species as pollen parent. The hypoaneuploid F_1 hybrids are highly heterozygous for all chromosomes except the one rendered monosomic, the loci of which are rendered hemizygous. Chromosome assignment using F_1 hypo-aneuploids is based on the "loss" of the *G. hirsutum* locus due to the missing chromosome transmitted by the *G. hirsutum* parent to the F_1. These lines have been used for chromosome assignment using PCR-based analyses for SSRs and SNPs (Liu, Saha et al. 2000; Hulse-Kemp, Ashrafi et al. 2014).

Several additional genomic resources for cotton have recently become available. A bacterial artificial chromosome (BAC) library resource has been developed to construct an initial physical map of the cultivated cotton *G. hirsutum* genome (Saski et. al. – personal communication). Based on the estimated genome size, 2.4Gb (Hendrix and Stewart 2005), the physical map provides ~85% genome coverage. The flanking clone ends of a proportion of the BACs were Sanger sequenced to provide a set of 179,209 BAC-end sequences (BESs). These BESs have recently been utilized along with whole genome resequencing data to develop hundreds of thousands of intra-specific (within cultivated *G. hirsutum* varieties) and inter-specific (between cultivated *G. hirsutum* and other species) SNPs (Hulse-Kemp et. al. – submitted).

While the physical map provides a template for localizing sequences for much of the cultivated cotton genome, the high levels of sequence redundancy associated with repetitive sequences and between homeologous regions of this recently formed allotetraploid make it difficult to accurately localize a significant proportion of genome sequences (Grover, Kim et al. 2007; Wang, Zhang et al. 2012). Recent developments of fluorescent *in situ* hybridization (FISH) methods have been reported for subgenome localization of BACs (Liu et al. 2015 – submitted). These FISH-based methods are effective but are still difficult and low-throughput. Thus accurate high-throughput methods for sequence localization for allotetraploid cotton are needed.

In this study, we develop a high-throughput method for localization of sequences using SNP genotyping for BAC-derived SNPs using whole-genome resequencing of two interspecific hypo-aneuploid F_1 cotton plants, one monosomic for A-subgenome chromosome 12 and the other for its D-subgenome homeolog, chromosome 26. Our localization results are validated by BAC-FISH subgenome assignments and linkage mapping of BES-derived SNPs. Analysis of BES-derived SNPs can then be used to enhance the recently reported allotetraploid cotton physical map for chromosomes 12 and 26. The sequence data provided here can be used by other researchers to localize sequences of interest to allotetraploid chromosomes 12 and 26.

Materials and Methods

Aneuploid sequencing

Two *Gossypium hirsutum* monosomic plants isogenic to cultivar Texas Marker-1 (TM-1) identified as missing a single chromosome, one missing chromosome 12 and one missing a chromosome 26, were each utilized as a female in a cross with *G. barbadense* doubled haploid line 3-79 to produce F_1 progeny. From each cross, an individual carrying the maternal deficiency was identified. The respective interspecific F_1 hypo-aneuploids were lacking *G. hirsutum* chromosome 12 (H12 - Lab ID: 199508048.03) and chromosome 26 (H26 - Lab ID: 201208076.04), and each was thus monosomic for the corresponding *G. barbadense* chromosome. The aneuploids were maintained under greenhouse conditions. Young leaf tissue was sampled from each plant and used to isolate genomic DNA using the Machrey-Nagel Plant Nucleo-spin kit following manufacturer instructions. Quality of DNA was assessed by running on a 1% agarose gel and then quantified with a PicoGreen® assay on the Synergy HT plate reader (Bio-Tek). Libraries were prepared from 1.5 µg of randomly sheared DNA from a Bioruptor instrument (Diagenode) for seven cycles of 15 seconds on and 90 seconds off with a short spin to gather the entire sample between cycles 4 and 5. Results of shearing were visualized by gel electrophoresis and then size selection was performed using AMPure XP beads to 300-500bp. NEBNext End Repair Module was used for end repair of fragments and then purified using AMPure beads. Adenine-addition and adapter ligations were performed and the final products were purified with AMPure beads. Enrichment PCR was performed for 14 cycles and then the PCR product was analyzed on a 1.5% agarose gel to confirm enrichment and size range of products. AMPure beads were used for a final round of purification. Final libraries were run on the Bioanlyzer (Agilent) to determine final library size and concentration. The two samples were sequenced on two Illumina HiSeq2500 lanes for paired-end sequencing (2x100bp). Raw, paired-read sequence files were uploaded to NCBI. Raw reads for euploid parents of F_1 hypo-aneuploid lines were obtained from the NCBI Small Read Archive under numbers SRX667500 (*G. hirsutum* cv. TM-1) and SRX669474 (*G. barbadense* line 3-79). TM-1 derived BAC-end sequences were also obtained from NCBI (LIBGSS_039228).

BAC-derived SNP genotyping

Raw sequence files for H12, H26, and parent samples were initially assessed for quality using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were then trimmed for low quality, adapters and any reads fewer than 40 bases were removed using the FastX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). Sequences for forward and reverse reads were then concatenated into a single file to be utilized as single-end read data. Read files were imported and aligned to BAC-end sequence references as in Hulse-Kemp et al. (submitted), and SNP positions were called using CLC Geomics Workbench v 6.0.2 (Valencia, USA). Variant files were filtered using the *in silico*-derived TM-1 homeo-SNPs, or inter-locus variants between homeologous regions in different subgenomes of the allotetraploid, as in Hulse-Kemp et al. (submitted). Variant files were further filtered to retain only homozygous loci, which were considered to be homozygous if all reads or all but one read for the locus represented the homozygous genotype.

Sequence localization

The number of homozygous variants identified for each BAC, including variants identified on both forward and reverse BAC-end sequences, was calculated for the H12, H26 and 3-79 samples. The proportion of homozygous variants in H12 and H26 for each BAC compared to the number identified with 3-79 was calculated. The results for H12 and H26 were each filtered to determine BACs for which over 75% of the homozygous variants found in 3-79 were identified as homozygous-like (mono-allelic) in a hypo-aneuploid. The results were further filtered to remove any BACs that were found in the opposite hypo-aneuploid sample, leading to final lists of BACs putatively localized to chromosomes 12 and 26. The placement of BACs in the cotton physical map v1 (Saski et al. – submitted) were identified for the putatively localized BACs. The BACs located in unlocalized contigs, singletons, or putatively placed in opposite subgenomes in the cotton physical map v1 (Saski et al. – submitted) were targeted for validation of placement by FISH and/or linkage mapping.

Linkage mapping

SNP and flanking sequences for the selected BACs were subjected to primer design using BatchPrimer3 for development of KASP assays (LGC Genomics). Primers were designed for 96 KASP assays using the following parameters: allele-specific primers and allele-flanking primers, Tm - optimum: 57ºC, minimum: 55ºC, maximum: 60ºC, max difference: 2ºC, product size- minimum: 50bp, optimum: 50bp, maximum: 100bp. Primers were synthesized by IDT and diluted according to KASP developer (LGC Genomics) instructions. Assays were used to genotype 118 F_2 (TM-1 x 3-79) individuals, two parents (*G. hirsutum* cultivar TM-1 and *G. barbadense* line 3-79), H12, H26 and F_1 (3-79 x TM-1) using the Fluidigm system in 96.96 dynamic array format. Genotypes for the markers genotyped here along with genotypes for the same F_2 individuals for 90 markers genotypes as determined in Hulse-Kemp et al. (submitted) were imported into JoinMap 4.1 (Van Ooijen 2006). Identical markers were removed and linkage mapping was performed using the maximum likelihood algorithm and Haldane's mapping function with default parameters. Linkage groups with LOD scores 5.0 or greater were retained.

Fluorescent in situ hybridization localization

As a complementary validation method, FISH on mitotic metaphase chromosomes was employed for subgenomic assignments of 40 BACs. Root tips from germinated seeds of *G. hirsutum* inbred TM-1 were utilized to produce metaphase chromosome preparations. Chromosome preparation was as described by Liu et al. (in preparation) using the smear method. DNA sequences of BAC clones were directly amplified by rolling circle amplification reactions and labelled by standard nick translation reactions. Repetitive DNA sequences were blocked by TM-1 Cot-1 DNA prepared according to (Zwick, Hanson et al. 1997). FISH and microscopy were performed as in Liu et al. (in preparation).

Results

Aneuploid sequencing, genotyping, and localization

Approximately 820 million reads (when considering forward and reverse reads as separate single reads) were produced on the HiSeq 2500 for H12 and H26 which is ~34.5X genome coverage for each genotype, assuming a genome size of 2.4Gb (Hendrix and Stewart 2005) (**Table 5.1**). After quality control and trimming, 6.60% and 6.88% of the reads were stringently mapped back to the BES reference for H12 and H26, respectively. When genotypes with completely homozygous read coverage were considered, H12 identified 12,827 homozygous SNP positions and H26 identified 10,314 homozygous SNP positions. These homozygous SNPs were considered to identify BACs that had 75% or higher homozygosity rates of the total SNPs identified on the respective BES(s). A total of 1,372 BACs were thereby identified with H12 and a total of 1,047 BACs were identified with H26.

Table 5.1 Sequencing and mapping statistics for analyzed euploid and hypo-aneuploid cotton samples.

Type/Species	Sample	Raw Reads (#)	Trimmed Reads (#)	Trimmed Reads (%)	Mapped Reads (#)	Mapped Reads (%)	Fraction of Reference Covered
Gossypium hirsutum L.	TM-1	816,832,880	796,773,046	97.54%	59,000,908	7.40%	84%
Gossypium barbadense L.	3-79	826,774,068	805,308,483	97.40%	49,280,456	6.12%	77%
F1 hypo-	H12	832,485,094	812,755,721	97.63%	53,634,336	6.60%	83%
aneuploid	H26	828,486,406	807,458,208	97.46%	55,583,602	6.88%	84%

The BACs localized to chromosomes 12 and 26 were used to identify the chromosomal association of the respective contigs of the cotton physical map (v1) by Saski et al. (submitted) (**Table 5.2**). The current physical map contains 85 contigs in the chromosome-26 minimum tiling path (MTP) and 173 contigs in the chromosome-12 MTP. Of these, this localization using H26 and H12 sequence was able to identify a total of 78 (91.8%) and 108 (62.4%) of the contigs in the D-subgenome and A-subgenome maps, respectively. Of the identified contigs from the current minimum tiling paths, 77.7% and 54.3% were identified by H26 and H12 as being placed correctly. Of the total 210 contigs identified with H26, 171 (81.5%) were located either on the correct chromosome or on unplaced contigs in the current (v1) cotton physical map. Of the total 234 contigs identified with H12, 182 (77.8%) were located either on the correct chromosome or on unplaced contigs in the current cotton physical map. In addition, 356 and 1,012 singleton BACs were also identified and localized using H26 and H12.

Table 5.2 Chromosomal placement of contigs and singleton BACs localized using interspecific F_1 hypo-aneuploid resequencing of H26 and H12 samples.

		Localized with H26		**Localized with H12**	
G. raimondii **Chromosome**	**Corresponding Allotetraploid Chromosomes (D & A)**	**Correct (D) Subgenome**	**Incorrect (A) Subgenome**	**Correct (A) Subgenome**	**Incorrect (D) Subgenome**
Chr01	16 & 7	0	0	1	6
Chr02	15 & 1	1	1	1	8
Chr03	17 & 2/3	1	0	2	1
Chr04	24 & 8	5	13	0	2
Chr05	14 & 2/3	2	0	1	6
Chr06	23 & 9	0	0	0	3
Chr07	21 & 11	2	1	0	0
Chr08	26 & 12	66 (30.5%)	14 (6.7%)	94 (40.2%)	12 (5.1%)
Chr09	19 & 4/5	5	3	3	4
Chr10	25 & 6	1	0	0	1
Chr11	20 & 10	0	0	2	1
Chr12	22 & 4/5	2	0	4	3
Chr13	18 & 13	0	2	1	2
Unplaced Contigs	-	64 (30.5%)	27 (12.9%)	57 (24.4%)	19 (8.1%)
Placed Singletons	-	9	68	67	1
Unplaced Singletons	-	279		944	
Total Contigs	-	**210**		**234**	
Total Singletons	-	**356**		**1012**	

Linkage mapping

A total of 78 of the 96 attempted SNP assays produced genotypes for the 118 F_2 individuals, i.e., an 81.25% success rate (**Table 5.3**). This success rate is similar to what was found in Hulse-Kemp et al. (submitted). The genotypes for these markers were combined with the genotypes for 88 markers obtained in Hulse-Kemp et al. (submitted) to produce linkage groups for chromosomes 12 and 26.

Compared to the linkage groups from the Hulse-Kemp et al (submitted), the linkage group for chromosome 12 was not greatly increased in size, as the majority of the additional markers were added to the central portion of the linkage group. However, the size of the linkage group for chromosome 26 was increased almost two-fold based on the addition of markers outside of the previous start and end markers, specifically increasing from 69.9cM to 120.6cM. When additional markers were included to expand the linkage groups determined in Hulse-Kemp et al. (submitted), the syntenic relationships of the markers were maintained (**Figure 5.1**).

Table 5.3 KASP assay results for markers used for linkage mapping of chromosomes 12 and 26.

Marker	Result	Marker Type	Linkage Map Placement
GH_TBb002E06f93	GOOD	Codominant	Chr12
GH_TBb004H02f447	GOOD	Codominant	Chr12
GH_TBb005G05r124	BAD	-	-
GH_TBb012C15r567	BAD	-	-
GH_TBb013C08f626	BAD	-	-
GH_TBb017F07f168	GOOD	Codominant	Chr26
GH_TBb017G09f668	GOOD	Codominant	Chr26

Table 5.3 (continued)

Marker	Result	Marker Type	Linkage Map Placement
GH_TBb017K02f288	GOOD	Codominant	Chr26
GH_TBb019I23r380	GOOD	Codominant	Chr26
GH_TBb020M04f512	GOOD	Codominant	Chr12
GH_TBb027J12f190	GOOD	Codominant	Chr26
GH_TBb027N23f106	GOOD	Codominant	LG4
GH_TBb028A01f104	BAD	-	-
GH_TBb028H07f252	BAD	-	-
GH_TBb030D18r167	GOOD	Codominant	Chr26
GH_TBb030F23f606	GOOD	Codominant	Chr12
GH_TBb032I08r364	GOOD	Codominant	Chr12
GH_TBb037O02f409	GOOD	Codominant	Chr12
GH_TBb038L02r120	GOOD	Codominant	Chr26
GH_TBb039G10r116	GOOD	Dominant	Chr12
GH_TBb039J18r363	BAD	-	-
GH_TBb041N20f148	GOOD	Codominant	Chr12
GH_TBb045J23f665	GOOD	Codominant	Chr26
GH_TBb047L01r545	GOOD	Codominant	Chr26
GH_TBb053G20r215	GOOD	Codominant	Chr12
GH_TBb055D14f77	GOOD	Codominant	Chr12
GH_TBb055G01f593	GOOD	Codominant	Chr12
GH_TBb056E20f284	GOOD	Dominant	Chr12
GH_TBb058D02r584	GOOD	Codominant	Chr12
GH_TBb058J22r97	BAD	-	-
GH_TBb064D14f440	GOOD	Codominant	Chr26
GH_TBb067B02r284	BAD	-	-
GH_TBb071D17r45	GOOD	Codominant	Chr26
GH_TBb076C12f116	GOOD	Dominant	Chr12
GH_TBb076J11r207	GOOD	Codominant	Chr12
GH_TBb076N02f211	GOOD	Codominant	Chr12
GH_TBb081N03r335	GOOD	Dominant	Chr12
GH_TBb082O05r676	GOOD	Codominant	Chr26
GH_TBb084I04f56	GOOD	Codominant	Chr26
GH_TBb084N02f154	GOOD	Codominant	Chr26
GH_TBb088J04r618	BAD	-	-
GH_TBb091H18f602	GOOD	Codominant	Chr26

Table 5.3 (continued).

Marker	Result	Marker Type	Linkage Map Placement
GH_TBb092H24r843	GOOD	Codominant	Chr12
GH_TBb093E14r388	GOOD	Codominant	Chr12
GH_TBb093P05r386	BAD	-	-
GH_TBb094C17r548	GOOD	Codominant	Chr12
GH_TBb101N03r254	GOOD	Codominant	Singleton
GH_TBb102I06r329	BAD	-	-
GH_TBb106G22f347	GOOD	Dominant	Chr26
GH_TBb112D19r714	BAD	-	-
GH_TBb112P02f406	GOOD	Codominant	Chr26
GH_TBb114J17f276	BAD	-	-
GH_TBb114O14r174	GOOD	Codominant	LG4
GH_TBh004J06r654	GOOD	Codominant	Chr26
GH_TBh004N05r366	GOOD	Codominant	Chr12
GH_TBh006F05f269	GOOD	Codominant	Chr12
GH_TBh014B03r458	GOOD	Codominant	Chr26
GH_TBh016C13r229	GOOD	Dominant	Chr12
GH_TBh019G10f154	GOOD	Codominant	Chr26
GH_TBh019N04r687	GOOD	Codominant	Chr12
GH_TBh021G03f144	GOOD	Codominant	Chr26
GH_TBh024K24f193	GOOD	Codominant	Singleton
GH_TBh026D11r272	GOOD	Dominant	LG3
GH_TBh026E23f403	GOOD	Dominant	Chr26
GH_TBh027F02r260	GOOD	Codominant	Chr12
GH_TBh027I07r285	GOOD	Codominant	LG3
GH_TBh031K23f554	GOOD	Codominant	Chr26
GH_TBh034C10r37	GOOD	Codominant	Chr26
GH_TBh037A12r355	GOOD	Codominant	Chr26
GH_TBh042J02f149	GOOD	Dominant	Singleton
GH_TBh053H10r320	GOOD	Codominant	Chr12
GH_TBh055E02r283	GOOD	Codominant	Chr26
GH_TBh057B09r285	GOOD	Codominant	Chr12
GH_TBh057M15f363	GOOD	Codominant	Chr12
GH_TBh060M10r161	GOOD	Codominant	Chr12
GH_TBh060O04f563	BAD	-	-
GH_TBh067F22r110	GOOD	Codominant	Chr26

Table 5.3 (continued)

Marker	Result	Marker Type	Linkage Map Placement
GH_TBh071P01r611	BAD	-	-
GH_TBh072O23r209	GOOD	Codominant	Chr26
GH_TBh074K07f374	GOOD	Codominant	Chr26
GH_TBh080E20f71	GOOD	Codominant	Chr12
GH_TBh081J24r373	GOOD	Dominant	Singleton
GH_TBh084L01f443	GOOD	Codominant	Chr12
GH_TBh085K07f494	GOOD	Dominant	Chr12
GH_TBh086M04f575	BAD	-	-
GH_TBh091O04f569	GOOD	Codominant	Chr12
GH_TBh096C21f212	BAD	-	-
GH_TBh101I02f180	GOOD	Codominant	Chr12
GH_TBh101N19f631	GOOD	Codominant	Singleton
GH_TBh102M23r296	GOOD	Codominant	Chr26
GH_TBh106M11f47	GOOD	Codominant	Chr12
GH_TBh110D10r554	GOOD	Codominant	Chr26
GH_TBh111N23r161	GOOD	Dominant	Chr26
GH_TBh114H09f90	GOOD	Codominant	Chr26
GH_TBh114L07r639	GOOD	Codominant	Chr26
GH_TBr312D16f133	GOOD	Codominant	Chr12

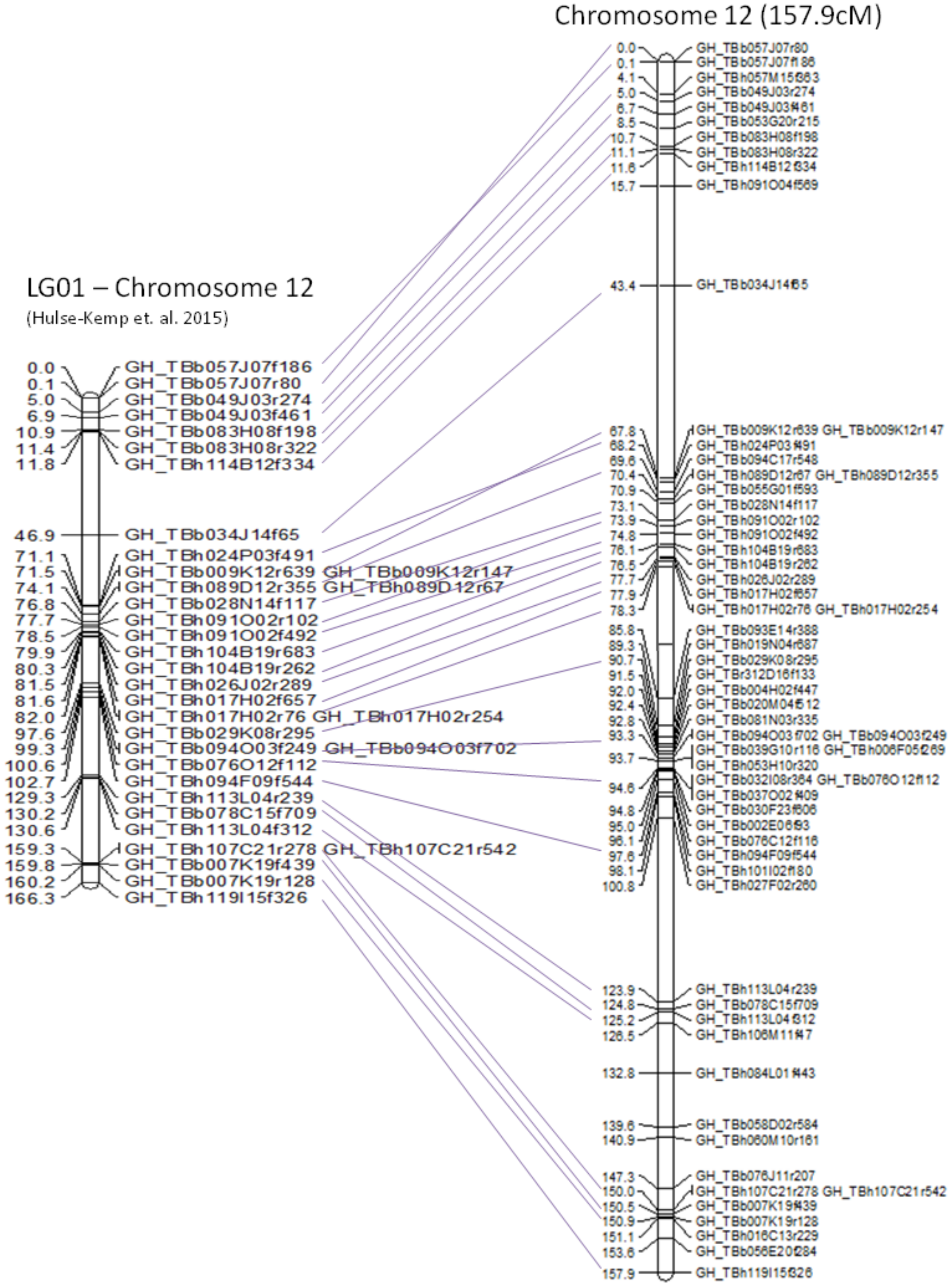


Figure 5.1 Expanded linkage maps for Chromosomes 12 and 26. Linkage maps were produced from 118 F2 individuals in JoinMap.

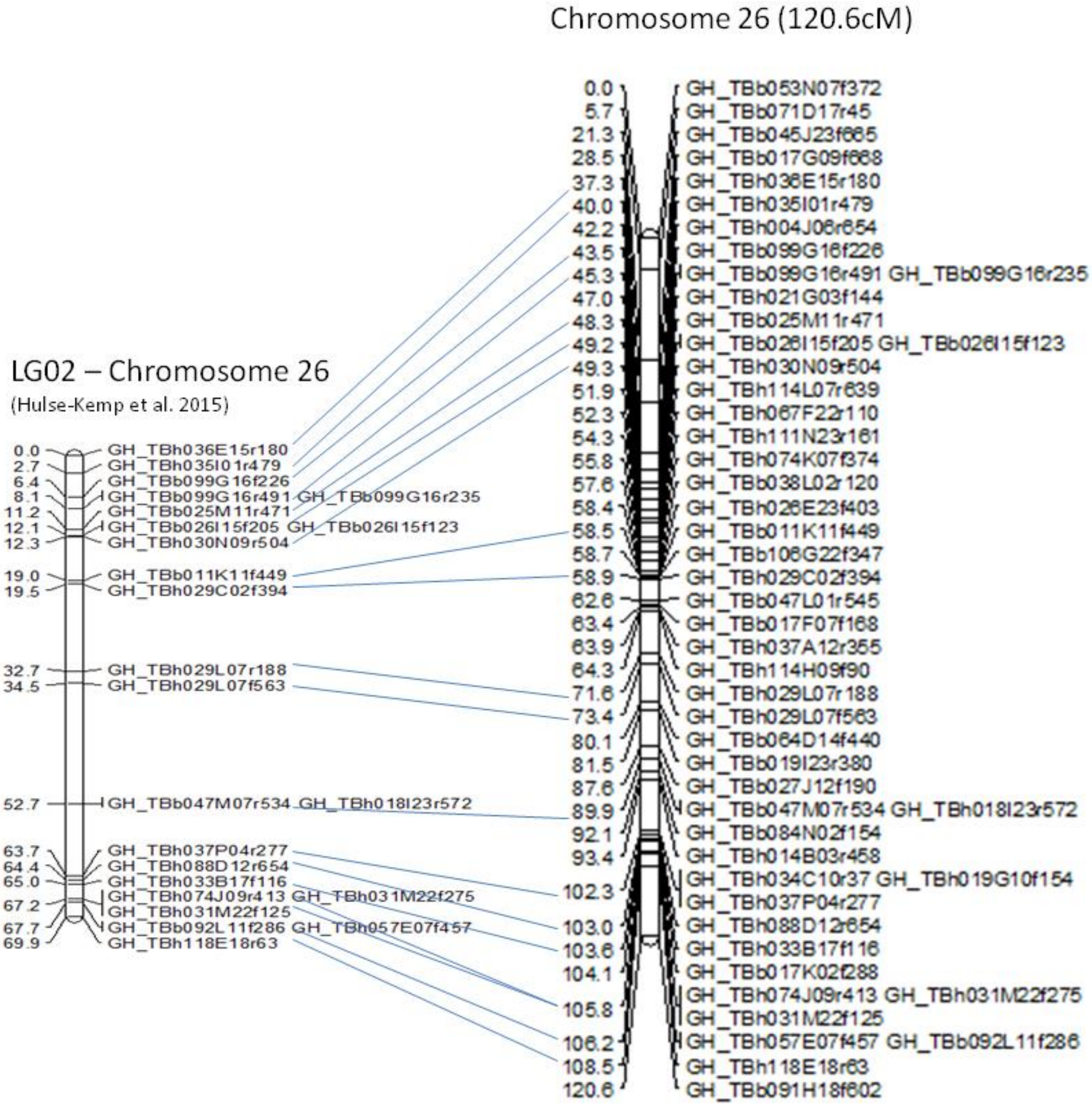


Figure 5.1 (continued)

Fluorescent in situ hybridization localization

A total of 40 BACs that had been identified using the F_1 hypo-aneuploid sequence were collaboratively FISHed, of those 19 BACs yielded site-specific fluorescent signals. Relative strengths and positions of the BAC-FISH signals among and within subgenomes and chromosomes allowed them and the associated BACs to be localized to a subgenome in the allotetraploid. Of the 19 BACs that were localized, 18 of them had been localized to the same subgenome using the H12/H26-based sequence localization method. An example for the localization of BAC GH_TBb071D17 (associated-SNP GH_TBb071D17r45) based on FISH results can be seen in **Figure 5.2**.

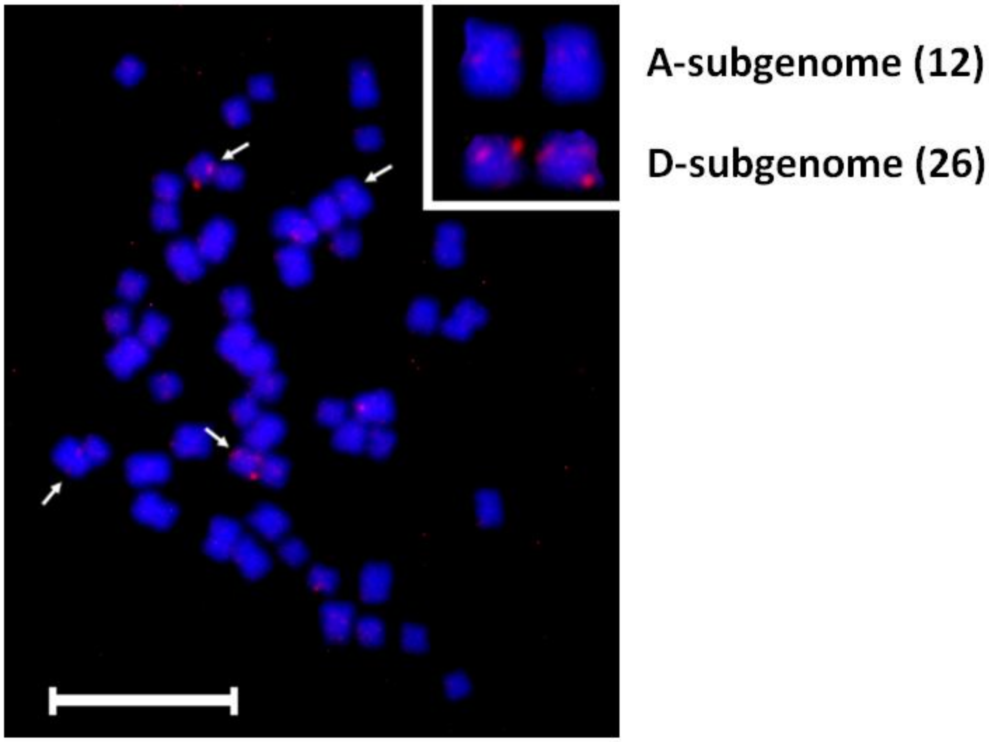


Figure 5.2 Subgenome localization by fluorescent *in situ* hybridization for BAC GH_TBb071D17. The arrows show the location of signals for GH_TBb071D17. The scale bar represents 10 micrometers.

Discussion

Efficacy and reliability of sequence localization

Resequencing of isogenic hypoaneuploid interspecific hybrids was found to be a cost-effective and highly efficient method of identifying and localizing large numbers of SNPs in the corresponding chromosomes. The strategy was assessed by two methods and the results of both assessments indicate that use of resequencing data of F_1 hypo-aneuploids for placement of BACs was highly reliable. The fluorescent *in situ* hybridization methods provided a straightforward visual assessment of the positions of BACs with associated SNPs (Liu et al. – submitted), while placement via linkage mapping with the BAC-associated SNPs determined corroborating placement. While it would be impossible to utilize FISH for placement of large numbers of BACs, this testing was able to show directly on the allotetraploid cotton chromosomes that localization with resequencing data was producing reliable results. Conversely, linkage mapping is readily applied to large numbers of SNPs, and thus complements the FISH-based approach well as a validation strategy. Taken together, these validation methods clearly indicate that resequencing can be utilized large scale for high-throughput localization of sequences. This resequencing-based method is particularly powerful in that the location information is being derived from the “loss” of locus information from the *G. hirsutum* parent, thus the localization is truly the location of the locus in cultivated cotton, *G. hirsutum*.

Advancements for cotton physical mapping

Cotton is like many other polyploid crops in that it has a very complex genome with large proportions of repetitive elements and homeologous regions. Due to its complex genome structure and the extensive similarity between large homeologous regions, sequence localization to chromosomes and subgenomes is difficult. One strategy for placement of BACs is to align at the contig level rather than individual BAC level to the reference genome, which is from an extant diploid relative. While the contig-based approach clearly improved placements, our results indicate that it is an imperfect solution. The majority of contigs identified from the Saski et al. physical map seem to be correctly localized, but *ca.* 20% of the contigs were localized to opposite homeologues in the current physical map version. These findings emphasize the difficulty of identifying the subgenomic associations of sequences in cotton and other complex polyploid crops, and also highlight the need for utilizing multiple validation methods. Such validation will be essential for developing an accurate reference genome for allotetraploid cotton.

The resequencing localization also allows for additional BACs to be incorporated into the physical map, as it has allowed for localization of a large number of unplaced contigs as well as many singleton BACs. Many of these contigs and singletons that have been localized here seem to be coming from peri-centromeric BACs, in that when the SNPs from these BACs are localized by linkage mapping, many are incorporated into the central regions of the linkage groups (**Figure 5.3**). These chromosome regions have large amounts of repetitive elements and low amounts of unique sequences, so most

BACs from these regions are very difficult to localize, and are not very amenable to fingerprinting for contig assembly or BAC alignment using BAC-end sequences. Therefore the resequencing method reported here seems to provide a way to identify and localize BACs that are recalcitrant to other methods of localization.

Benefits for complex genomes

The approaches taken here are applicable to many complex polyploid crops, as many of them have currently available aneuploid derived lines, such as in bread wheat, durum wheat, oat, and rye (Merker 1973; Joppa and Williams 1988; Oliver, Tinker et al. 2013; Khlestkina 2014). Utility of the interspecific F_1 hypo-aneuploids that were utilized in this study are somewhat customizable based on the line or species used to cross with the aneuploid line to develop the F_1 hypo-aneuploid, in that the cross lines can either be the same or a different species, as the case here. The choice of parent will determine the amount of polymorphism between the parental lines of the F_1, and allows the investigator to increase or decrease the amount of available diversity obtained in the sequence. For cultivated *G. hirsutum*, a species very low in diversity (Van Esbroeck and Bowman 1998), an interspecific cross with *G. barbadense* was chosen in order to create levels of diversity that could be identifiable within the span of a next-generation sequencing read, i.e. 100bp (Hulse-Kemp et al. – submitted). Generation of resequencing data from F_1 hypo-aneuploid lines would provide valuable public sequence resources, as the entire crop community can use them to localize sequences; these will collectively contribute towards development of one or more high-quality reference genomes.

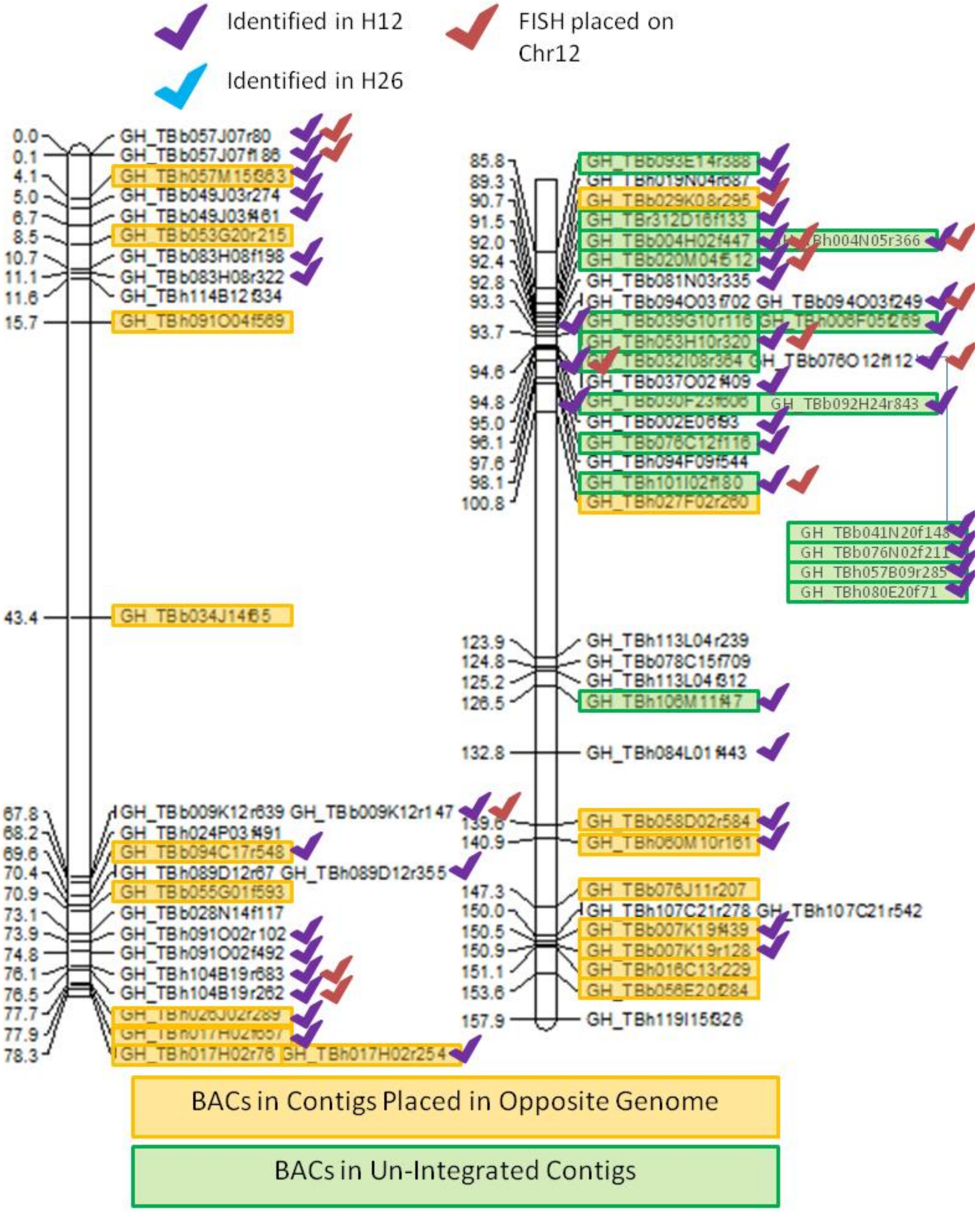


Figure 5.3 Chromosome 12 and 26 linkage groups depicting all localization information for linkage mapping, F1 hypo-aneuploid resequencing, and fluorescent *in situ* hybridization placement. Contigs which were unplaced in the cotton physical map v1 (Saski et al – submitted) are highlighted.

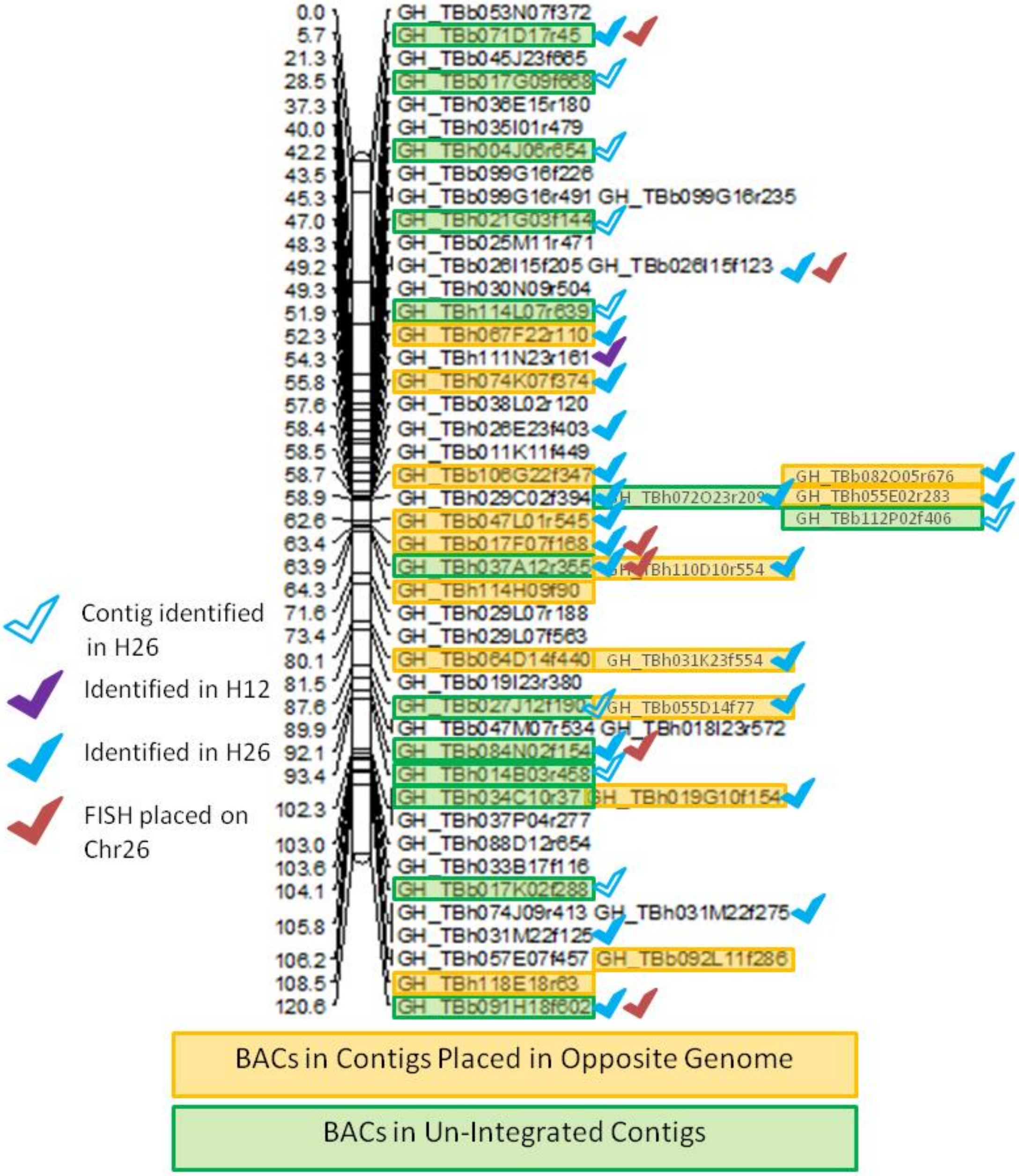


Figure 5.3 (continued)

CHAPTER VI

CONCLUSIONS

Using transcriptome sequencing, a large set of 62,832 SNPs relative to cultivated *G. hirsutum* were developed, which will allow for the first high-density mapping of genes from five wild species that affect traits of interest. Markers associated with functional differences between species are essential for generating a feasible system for germplasm introgression via marker-assisted breeding for beneficial agronomic and fiber characteristics. Future large-scale mapping, fine-mapping, and genome-wide association analysis efforts to associate markers developed here as diagnostic markers for traits of interest will allow for marker-assisted selection and backcrossing to speed up introgression efforts. Advancements in interspecific germplasm introgression are likely to create opportunities for profound improvement of cotton *G. hirsutum* cultivars.

With resequencing data from 14 cotton lines, the largest set of intraspecific and interspecific SNPs for cultivated cotton to date has been developed. These SNPs are associated with BACs and will serve as an interface between future physical and genetic maps. They were developed using genomic sequences from multiple lines and species aligned to BAC-end sequences generated by Sanger sequencing, which provided a high quality reference. Experimental validation was highly successful and indicated that the SNPs will allow for future high-density mapping. Furthermore, additional lines can be resequenced and quickly genotyped for the identified SNP positions using GATK software. The developed markers complement the developed genic-based SNPs and simple sequence repeat markers, and provide a largely evenly distributed set of markers for mapping the entire cotton genome at high-density. This will promote more extensive genomic-based studies and breeding of cotton.

Using the SNP data sets developed here and other available SNP data sets, a standardized high-throughput genotyping tool, the CottonSNP63K and accompanying cluster file was developed for the cotton community with limited ascertainment bias over a wide range of cotton germplasm. The array was used to produce the two highest density genetic maps for cotton to date, and the quality of the maps was demonstrated by synteny with the *G. raimondii* reference genome. The new array provides a useful resource for analyzing genome-wide variation in allotetraploid cotton. Once the allotetraploid cotton genome sequence is available, these maps will be a resource to validate and refine the cotton genome assemblies, as well as provide a resource to directly integrate physical and genetic resources. The array and maps provide a foundation for the genetic dissection of agronomically and economically important traits, and crop improvement through genomics-assisted selection. It will also foster positional cloning and genome assembly efforts.

Resequencing data from F_1 hypo-aneuploids was used in concert with BAC-end sequences to accurately localize sequence contigs to individual cotton chromosomes. As SNPs are widely distributed throughout the genome, this approach can be applied more widely to localize sequences in most areas of the genome. This will allow for enhancement as well as validation of high-quality reference sequences by providing an additional method for sequence localization and validation of placements for currently localized sequences. This approach will be important for crops with highly complex genomes, like cotton, that need multiple lines of validation.

The SNP data sets developed, the high-throughput genotyping array and the developed chromosomal localization method all provide additional validation and placement methods that can be directly integrated with the physical map constructed for cultivated cotton in order to ultimately produce a high-quality draft genome sequence for *Gossypium hirsutum* (L.).

REFERENCES

Altaf, M., J. Stewart, et al. (1997). Survey of cotton germplasm for terpenoid aldehydes important in host plant resistance. Special Reports-University of Arkansas Agricultural Experiment Station.

An, C., S. Saha, et al. (2008). "Cotton (*Gossypium* spp.) R2R3-MYB transcription factors SNP identification, phylogenomic characterization, chromosome localization, and linkage mapping." Theoretical and Applied Genetics **116**(7): 1015-1026.

Applequist, W. L., R. Cronn, et al. (2001). "Comparative development of fiber in wild and cultivated cotton." Evolution & Development **3**(1): 3-17.

Bayer, M., I. Milne, et al. (2011). "Comparative visualization of genetic and physical maps with Strudel." Bioinformatics **27**(9): 1307-1308.

Bell, A. A., A. Forest Robinson, et al. (2014). "Registration of LONREN-1 and LONREN-2 germplasm lines of Upland Cotton resistant to reniform nematode." Journal of Plant Registrations **8**(2): 187-190.

Bell, A. A., J. Quintana, et al. (2013). Pest Resistance and agronomic performance of advanced BARBREN lines compared to commercial cultivars and the germplasm line BARBREN 713. Beltwide Cotton Conference, San Antonio, TX, National Cotton Council of America.

Bianco, L., A. Cestaro, et al. (2014). "Development and validation of a 20K single nucleotide polymorphism (SNP) whole genome genotyping array for apple (Malus × domestica Borkh)." PLoS One **9**(10): e110377.

Blenda, A., D. D. Fang, et al. (2012). "A high density consensus genetic map of tetraploid cotton that integrates multiple component maps through molecular marker redundancy check." PLoS One **7**(9): e45739.

Bohra, A., A. Dubey, et al. (2011). "Analysis of BAC-end sequences (BESs) and development of BES-SSR markers for genetic mapping and hybrid purity assessment in pigeonpea (*Cajanus* spp.)." BMC Plant Biology **11**.

Briddon, R. and P. Markham (2000). "Cotton leaf curl virus disease." Virus research **71**(1): 151-159.

Brown, M. S. (1980). "Identification of the chromosomes of *Gossypium hirsutum* L. by means of translocations." Journal of Heredity **71**(4): 266-274.

Brubaker, C. L., A. H. D. Brown, et al. (1999). "Production of fertile hybrid germplasm with diploid Australian Gossypium species for cotton improvement." Euphytica **108**(3): 199-213.

Buriev, Z. T., S. Saha, et al. (2010). "Clustering, haplotype diversity and locations of MIC-3: a unique root-specific defense-related gene family in Upland cotton (*Gossypium hirsutum* L.)." Theoretical and Applied Genetics **120**(3): 587-606.

Byers, R. L., D. B. Harker, et al. (2012). "Development and mapping of SNP assays in allotetraploid cotton." Theoretical and Applied Genetics **124**(7): 1201-1214.

Chen, H., W. Xie, et al. (2013). "A high-density SNP genotyping array for rice biology and molecular breeding." Molecular Plant: sst135.

Chen, W., J. B. Yao, et al. (2014). "The development of specific SNP markers for chromosome 14 in cotton using next-generation sequencing." Plant Breeding **133**(2): 256-261.

Chen, Z. J., B. E. Scheffler, et al. (2007). "Toward sequencing cotton (Gossypium) genomes." Plant Physiology **145**(4): 1303-1310.

Ching, A., K. S. Caldwell, et al. (2002). "SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines." BMC Genetics **3**.

Dalton-Morgan, J., A. Hayward, et al. (2014). "A high-throughput SNP array in the amphidiploid species *Brassica napus* shows diversity in resistance genes." Functional & Integrative Genomics **14**(4): 643-655.

Desai, A., P. W. Chee, et al. (2006). "Chromosome structural changes in diploid and tetraploid A genomes of Gossypium." Genome / National Research Council Canada **49**(4): 336-345.

Elshire, R. J., J. C. Glaubitz, et al. (2011). "A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species." PLoS One **6**(5).

Fang, D. D., L. L. Hinze, et al. (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.

Frelichowski, J. E., M. B. Palmer, et al. (2006). "Cotton genome mapping with new microsatellites from Acala 'Maxxa' BAC-ends." Molecular Genetics and Genomics **275**(5): 479-491.

Ganal, M. W., G. Durstewitz, et al. (2011). "A large maize (*Zea mays* L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome." PLoS One **6**(12): e28334.

Ganal, M. W., A. Polley, et al. (2012). "Large SNP arrays for genotyping in crop plants." Journal of Biosciences **37**(5): 821-828.

Gao, W., Z. J. Chen, et al. (2006). "Wide-cross whole-genome radiation hybrid mapping of the cotton (*Gossypium barbadense* L.) genome." Molecular Genetics and Genomics : MGG **275**(2): 105-113.

Gao, W., Z. J. Chen, et al. (2004). "Wide-cross whole-genome radiation hybrid mapping of cotton (*Gossypium hirsutum* L.)." Genetics **167**(3): 1317-1329.

Goecks, J., A. Nekrutenko, et al. (2010). "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences." Genome Biology **11**(8): R86.

Gore, M. A., D. D. Fang, et al. (2014). "Linkage map construction and quantitative trait locus analysis of agronomic and fiber quality traits in cotton." The Plant Genome.

Gore, M. A., R. G. Percy, et al. (2012). "Registration of the TM-1/NM24016 cotton recombinant inbred mapping population." Journal of Plant Registrations **6**(1): 124-127.

Grover, C., X. Zhu, et al. (2014). "Molecular confirmation of species status for the allopolyploid cotton species, *Gossypium ekmanianum* Wittmack." Genetic Resources and Crop Evolution: 1-12.

Grover, C. E., H. Kim, et al. (2007). "Microcolinearity and genome evolution in the AdhA region of diploid and polyploid cotton (*Gossypium*)." The Plant Journal : for Cell and Molecular Biology **50**(6): 995-1006.

Guo, W., C. Cai, et al. (2007). "A microsatellite-based, gene-rich linkage map reveals genome structure, function and evolution in *Gossypium*." Genetics **176**(1): 527-541.

Hamilton, J. P., C. N. Hansey, et al. (2011). "Single nucleotide polymorphism discovery in elite North American potato germplasm." BMC Genomics **12**: 302.

Han, Y., D. Chagne, et al. (2009). "BAC-end sequence-based SNPs and bin mapping for rapid integration of physical and genetic maps in apple." Genomics **93**(3): 282-288.

Hansey, C. N., B. Vaillancourt, et al. (2012). "Maize (*Zea mays* L.) genome diversity as revealed by RNA-sequencing." PLoS One **7**(3): e33071.

Hawkins, J. S., H. Kim, et al. (2006). "Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*." Genome Research **16**(10): 1252-1261.

Hendrix, B. and J. M. Stewart (2005). "Estimation of the nuclear DNA content of *Gossypium* species." Annals of Botany **95**(5): 789-797.

Hinze, L. L., D. D. Fang, et al. (2015). "Molecular characterization of the *Gossypium* Diversity Reference Set of the US National Cotton Germplasm Collection." Theoretical and Applied Genetics **128**(2): 313-327.

Hulse-Kemp, A. M., H. Ashrafi, et al. (2014). "Development and bin mapping of gene-associated interspecific SNPs for cotton (*Gossypium hirsutum* L.) introgression breeding efforts." BMC Genomics **15**(1): 945.

Hyten, D. L., S. B. Cannon, et al. (2010). "High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence." BMC Genomics **11**.

Islam, M. S., G. N. Thyssen, et al. (2014). "Detection, validation and application of genotyping-by-sequencing based single nucleotide polymorphisms in upland cotton (*Gossypium hirsutum* L.)." The Plant Genome.

Jayaraj, S. and P. Palaniswamy (2005). Host plant resistance in cotton to major insect pests: Perspectives and progress. Sustainable Insect Pest Management. S. Ignacimuthu and S. Jayaraj. New Delhi, Alpha Science Int'l Ltd.

Jenkins, J. N., J. C. McCarty, et al. (2007). "Genetic effects of thirteen *Gossypium barbadense* L. chromosome substitution lines in topcrosses with Upland cotton cultivars: II. Fiber quality traits." Crop Science **47**(2): 561-570.

Jenkins, J. N., J. Wu, et al. (2006). "Genetic effects of thirteen *Gossypium barbadense* L. chromosome substitution lines in topcrosses with Upland cotton cultivars: I. Yield and yield components." Crop Science **46**(3): 1169-1178.

Joppa, L. and N. Williams (1988). "Langdon durum disomic substitution lines and aneuploid analysis in tetraploid wheat." Genome / National Research Council Canada = Genome / Conseil national de recherches Canada **30**(2): 222-228.

JW, V. A. N. O. (2011). "Multipoint maximum likelihood mapping in a full-sib family of an outbreeding species." Genetics Research **93**(5): 343-349.

Kaur, S., M. G. Francki, et al. (2012). "Identification, characterization and interpretation of single-nucleotide sequence variation in allopolyploid crop species." Plant Biotechnology Journal **10**(2): 125-138.

Keller, O., M. Kollmar, et al. (2011). "A novel hybrid gene prediction method employing protein multiple sequence alignments." Bioinformatics **27**(6): 757-763.

Khlestkina, E. K. (2014). "Current applications of wheat and wheat–alien precise genetic stocks." Molecular Breeding **34**(2): 273-281.

Kohel, R., T. Richmond, et al. (1970). "Texas marker-1. Description of a genetic standard for *Gossypium hirsutum* L." Crop Science **10**(6): 670-671.

Lacape, J. M., M. Claverie, et al. (2012). "Deep sequencing reveals differences in the transcriptional landscapes of fibers from two cultivated species of cotton." PLoS One **7**(11): e48855.

Lacape, J. M., T. B. Nguyen, et al. (2003). "A combined RFLP-SSR-AFLP map of tetraploid cotton based on a *Gossypium hirsutum* x *Gossypium barbadense* backcross population." Genome **46**(4): 612-626.

Lam, H. M., X. Xu, et al. (2010). "Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection." Nature Genetics **42**(12): 1053-U1041.

Lewin, H. A., D. M. Larkin, et al. (2009). "Every genome sequence needs a good map." Genome Research **19**(11): 1925-1928.

Li, F. G., G. Y. Fan, et al. (2014). "Genome sequence of the cultivated cotton *Gossypium arboreum*." Nature Genetics **46**(6): 567-572.

Li, H., B. Handsaker, et al. (2009). "The Sequence Alignment/Map format and SAMtools." Bioinformatics **25**(16): 2078-2079.

Liu, S., S. Saha, et al. (2000). "Chromosomal assignment of microsatellite loci in cotton." Journal of Heredity **91**(4): 326-332.

Lubbers, E. L., P. W. Chee, et al. (2003). Genetic diversity of U.S. upland cotton. Georgia Cotton Research and Extension Reports. C. Hester, University of Georgia**:** 120-130.

Luo, S., J. Mach, et al. (2012). "The cotton centromere contains a Ty3-gypsy-like LTR retroelement." PLoS One **7**(4): e35261.

Mammadov, J., R. Aggarwal, et al. (2012). "SNP markers and their impact on plant breeding." International Journal of Plant Genomics **2012**.

Manly, K. F., R. H. Cudmore, Jr., et al. (2001). "Map Manager QTX, cross-platform software for genetic mapping." Mammalian Genome : Official Journal of the International Mammalian Genome Society **12**(12): 930-932.

Matvienko, M., A. Kozik, et al. (2013). "Consequences of normalizing transcriptomic and genomic libraries of plant genomes using a duplex-specific nuclease and tetramethylammonium chloride." PLoS One **8**(2): e55913.

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

Merker, A. (1973). "Identification of aneuploids in a line of hexaploid Triticale." Hereditas **74**(1): 1-6.

Oliver, R. E., N. A. Tinker, et al. (2013). "SNP discovery and chromosome anchoring provide the first physically-anchored hexaploid oat map and reveal synteny with model species." PLoS One **8**(3): e58068.

Ollitrault, P., J. Terol, et al. (2012). "SNP mining in *C. clementina* BAC end sequences; transferability in the Citrus genus (Rutaceae), phylogenetic inferences and perspectives for genetic mapping." BMC Genomics **13**.

Page, J. T., A. R. Gingle, et al. (2013). "PolyCat: a resource for genome categorization of sequencing reads from allopolyploid organisms." G3 **3**(3): 517-525.

Page, J. T., M. D. Huynh, et al. (2013). "Insights into the evolution of cotton diploids and polyploids from whole-genome re-sequencing." G3: Genes| Genomes| Genetics **3**(10): 1809-1818.

Paterson, A. H., J. F. Wendel, et al. (2012). "Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres." Nature **492**(7429): 423-427.

Phillips, L. and M. Strickland (1966). "The cytology of a hybrid between *Gossypium hirsutum* and *G. longicalyx*." Canadian Journal of Genetics and Cytology **8**(1): 91-95.

Poland, J. A., P. J. Brown, et al. (2012). "Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach." PLoS One **7**(2).

R Development Core Team (2010). R: A language and environment for statistical computing. R. D. C. Team. Vienna, R Foundation for Statistical Computing.

Rai, K. M., S. K. Singh, et al. (2013). "Large-scale resource development in *Gossypium hirsutum* L. by 454 sequencing of genic-enriched libraries from six diverse genotypes." Plant Biotechnology Journal **11**(8): 953-963.

Reinisch, A. J., J.-M. Dong, et al. (1994). "A detailed RFLP map of cotton, *Gossypium hirsutum* x *Gossypium barbadense*: chromosome organization and evolution in a disomic polyploid genome." Genetics **138**(3): 829-847.

Richmond, T. (1950). "Procedures and methods of cotton breeding with special reference to American cultivated species." Advances in Genetics **4**: 213-245.

Robinson, A., A. Bell, et al. (2007). "Introgression of resistance to nematode into Upland cotton (*Gossypium hirsutum*) from *Gossypium longicalyx*." Crop Science **47**(5): 1865-1877.

Robinson, A. F., A. A. Bell, et al. (2007). "Introgression of resistance to nematode *Rotylenchulus reniformis* into upland cotton (*Gossypium hirsutum*) from *Gossypium longicalyx*." Crop Science **47**(5): 1865-1877.

Rong, J., C. A. Abbey, et al. (2004). "A 3347-locus genetic recombination map of sequence-tagged sites reveals features of genome organization, transmission and evolution of cotton (Gossypium)." Genetics **166**(1): 389-417.

Rong, J., J. E. Bowers, et al. (2005). "Comparative genomics of Gossypium and Arabidopsis: unraveling the consequences of both ancient and recent polyploidy." Genome Research **15**(9): 1198-1210.

Saha, S., D. A. Raska, et al. (2006). "Upland cotton (*Gossypium hirsutum* L.) x Hawaiian cotton (*G. tomentosum* Nutt. ex. Seem) F1 hybrid hypoaneuploid chromosome substitution series." Journal of Cotton Science **10**: 146-154.

Saha, S., D. Stelly, et al. (2012). "Chromosome substitution lines: concept, development and utilization in the genetic improvement of Upland cotton." Plant Breeding. InTech, Rijeka, Croatia.: 107-128.

Saha, S., J. Wu, et al. (2004). "Effect of chromosome substitutions from *Gossypium barbadense* L. 3-79 into *G. hirsutum* L. TM-1 on agronomic and fiber traits." Journal of Cotton Science **8**: 162-169.

Saha, S., J. Wu, et al. (2013). "Interspecific chromosomal effects on agronomic traits in *Gossypium hirsutum* by AD analysis using intermated *G. barbadense* chromosome substitution lines." Theoretical and Applied Genetics **126**(1): 109-117.

Shappley, Z. W., J. N. Jenkins, et al. (1998). "An RFLP linkage map of Upland cotton, *Gossypium hirsutum* L." Theoretical and Applied Genetics **97**(5-6): 756-761.

Shen, X. L., W. Guo, et al. (2007). "Genetic mapping of quantitative trait loci for fiber quality and yield trait by RIL approach in Upland cotton." Euphytica **155**(3): 371-380.

Shen, Y. J., H. Jiang, et al. (2004). "Development of genome-wide DNA polymorphism database for map-based cloning of rice genes." Plant Physiology **135**(3): 1198-1205.

Shi, Y., W. Li, et al. (2014). "Constructing a high-density linkage map for *Gossypium hirsutum* × *Gossypium barbadense* and identifying QTLs for lint percentage." Journal of Integrative Plant Biology.

Song, Q., D. L. Hyten, et al. (2013). "Development and evaluation of SoySNP50K, a high-density genotyping array for soybean." PLoS One **8**(1): e54985.

Stelly, D., S. Saha, et al. (2005). "Registration of 17 Upland (*Gossypium hirsutum*) cotton germplasm lines disomic for different *G. barbadense* chromosome or arm substitutions." Crop Science **45**(6): 2663-2665.

Tang, S., Z. Teng, et al. (2014). "Construction of genetic map and QTL analysis of fiber quality traits for Upland cotton (*Gossypium hirsutum* L.)." Euphytica: 1-19.

Tinker, N. A., S. Chao, et al. (2014). "A SNP genotyping array for hexaploid oat." The Plant Genome **7**(3).

Truco, M. J., H. Ashrafi, et al. (2013). "An ultra-high-density, transcript-based, genetic map of lettuce." G3: Genes| Genomes| Genetics **3**(4): 617-631.

Tyagi, P., M. A. Gore, et al. (2014). "Genetic diversity and population structure in the US Upland cotton (*Gossypium hirsutum* L.)." Theoretical and Applied Genetics **127**(2): 283-295.

Ulloa, M., I. Y. Abdurakhmonov, et al. (2013). "Genetic diversity and population structure of cotton (*Gossypium* spp.) of the New World assessed by SSR markers." Botany **91**(4): 251-259.

Van Deynze, A., K. Stoffel, et al. (2009). "Sampling nucleotide diversity in cotton." BMC plant biology **9**: 125.

Van Esbroeck, G. and D. T. Bowman (1998). "Cotton germplasm diversity and its importance to cultivar development." Journal of Cotton Science **2**(3): 121-129.

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

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

Wang, H., R. V. Penmetsa, et al. (2012). "Development and characterization of BAC-end sequence derived SSRs, and their incorporation into a new higher density genetic map for cultivated peanut (*Arachis hypogaea* L.)." BMC Plant Biology **12**: 10.

Wang, K., X. Song, et al. (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, K., Z. Wang, et al. (2012). "The draft genome of a diploid cotton *Gossypium raimondii*." Nature Genetics **44**(10): 1098-1103.

Wang, K., W. P. Zhang, et al. (2012). "Localization of high level of sequence conservation and divergence regions in cotton." Theoretical and Applied Genetics **124**(7): 1173-1182.

Wang, S., D. Wong, et al. (2014). "Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array." Plant Biotechnology Journal.

Weaver, D. B., R. B. Sikkens, et al. (2013). "RENlon and its effects on agronomic and fiber quality traits in upland cotton." Crop Science **53**(3): 913-920.

Wendel, J. F., C. Brubaker, et al. (2009). Evolution and natural history of the cotton genus. Genetics and Genomics of Cotton. A. H. Paterson. New York, Springer. **3:** 3-22.

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

Yan, J., T. Shah, et al. (2009). "Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers." PLoS One **4**(12): e8451.

Yik, C.-P. and W. Birchfield (1984). "Resistant germplasm in *Gossypium* species and related plants to *Rotylenchulus reniformis*." Journal of Nematology **16**(2): 146-153.

Yu, J., S. Jung, et al. (2014). "CottonGen: a genomics, genetics and breeding database for cotton research." Nucleic Acids Research **42**(Database issue): D1229-1236.

Yu, J. Z., R. J. Kohel, et al. (2012). "A high-density simple sequence repeat and single nucleotide polymorphism genetic map of the tetraploid cotton genome." G3 **2**(1): 43-58.

Zhang, K., J. Zhang, et al. (2012). "Genetic mapping and quantitative trait locus analysis of fiber quality traits using a three-parent composite population in upland cotton (*Gossypium hirsutum* L.)." Molecular Breeding **29**(2): 335-348.

Zhang, Z., M. Hu, et al. (2009). "Construction of a comprehensive PCR-based marker linkage map and QTL mapping for fiber quality traits in upland cotton (*Gossypium hirsutum* L.)." Molecular Breeding **24**(1): 49-61.

Zheng, X. (2012). High-Resolution Recombination to Dissect an Alien Segment of Cotton Chromosome-11 with Resistance to Reniform Nematodes. Plant and Animal Genome XX Conference, San Deigo.

Zheng, X., D. Levine, et al. (2012). "A high-performance computing toolset for relatedness and principal component analysis of SNP data." Bioinformatics **28**(24): 3326-3328.

Zhu, Q. H., A. Spriggs, et al. (2014). "Transcriptome and complexity-reduced, DNA-based identification of intraspecies single-nucleotide polymorphisms in the polyploid *Gossypium hirsutum* L." G3: Genes| Genomes| Genetics **4**(10): 1893-1905.

Zhulidov, P. A., E. A. Bogdanova, et al. (2004). "Simple cDNA normalization using kamchatka crab duplex-specific nuclease." Nucleic Acids Research **32**(3): e37-e37.

Zwick, M. S., R. E. Hanson, et al. (1997). "A rapid procedure for the isolation of Cot-1 DNA from plants." Genome / National Research Council Canada **40**(1): 138-142.