IDENTIFYING, MAPPING AND OVERCOMING GENOMIC IMPEDIMENTS TO INTRASPECIFIC GENETIC IMPROVEMENT OF UPLAND COTTON THROUGH INTERSPECIFIC HYBRIDIZATION AND INTROGRESSION

A Dissertation

by

LUIS MIGUEL DE SANTIAGO

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
Co-Chair of Committee,	Clare A. Gill
Committee Members,	Seth C. Murray
	Alan E. Pepper
	Mauricio Ulloa
Chair of Intercollegiate Faculty,	David W. Threadgill

August 2020

Major Subject: Genetics

ABSTRACT

Cotton, the leading natural textile fiber, develops as modified seed trichomes, produced in copious amounts by genetically elite cultivars of two domesticated New World Upland tetraploid (2n=4x=52) species, Upland cotton (*Gossypium hirsutum* L.) and Pima cotton (*G. barbadense* L.). Future cultivars must improve production sustainability, economic yield, and address the many challenges caused by global climate change. Their creation will require beneficial genetic diversity, relevant recombination, breeding, and selection. This work was undertaken [1] to better understand available diversity and natural constraints on its use, especially recombination, [2] to increase genetic diversity, and [3] to improve recombination as a tool for extracting value from diversity.

Aside from transgenes and occasional mutations, induced or created, new cotton cultivars are created from fast-track breeding programs that rely exclusively on genetic variation among cultivars and elite breeding germplasm, not on wild intraspecific or interspecific accessions; cyclic re-use of elite germplasm is implicit. Using CottonSNP63K data and a long-read genomic assembly, I characterized the genomic distribution of genetic diversity by identifying, mapping, and comparing haplotypic structures and their diversity among 257 elite Upland lines and 71 non-elite *G. hirsutum* accessions. Independent analyses using comparable types of data for 9 at intra- and interspecific linkage mapping populations revealed their recombination patterns. Comparisons showed strong relationships between static haplotypic blocks and low experimental recombination rates. Genomic characterization of highly and lowly recombinant regions revealed significant associations and correlations between recombination and transposable element densities and biological pathways between three allotetraploid cotton species and *G. hirsutum*.

An intriguing observation during linkage analyses was that hybridization with wild allotetraploid cotton species led to novel recombination events that disrupted long-standing haplotypic blocks. Such disruption and the recovery of novel genetic variation suggests potential for a new era in Upland cotton improvement involving extensive use interspecific hybridization to disrupt haplotypic blocks, as well as diversify genes. Indeed, this was demonstrated through recombination modeling and through the development of a chromosome segment substitution lines consisting of introgressed *G. tomentosum* chromatin.

These findings indicate potentially major if not revolutionary ramifications for breeding Upland cotton, and likely Pima, as well.

DEDICATION

To my parents Delia and Juan De Santiago, for demonstrating dedication and perseverance in the face of adversity. To my sister Maria, who continues to be the foundation of my family allowing my pursuits to be possible. Lastly, to old and new friends who have never failed to provide me with spirit.

ACKNOWLEDGEMENTS

A special acknowledgement and appreciation to my advisory, Dr. David Stelly, for affording me the opportunity to take on this research project and for the mentorship he has provided along the way. Thank you for allowing me to explore new ideas while keeping me grounded and for introducing me to numerous collaborators. Your positive guidance will have a long-lasting impact on my life. I also want to thank my dissertation committee members who have given me the honor of allowing me to be their pupil: Dr. Clare Gill, Dr. Seth Murray, Dr. Alan Pepper, and Dr. Mauricio Ulloa. This dissertation would not have been possible without their support. A special thanks to Dr. Jefferey Chen for not only providing funds for my research, but for giving me the opportunity to be a part of a great research team.

I am very grateful for my lab mates, past and present, for helping me both professionally and personally along the way: Yu-Ming Lin, Ammani Kyanam, Bree Vculek, Christian Hitzelberger, Jamshaid Junaid, Kübra Velioğlu, and Dr. Robert Vaughn. Their friendship and guidance have made this journey a much more meaningful experience. I also want to thank all the student workers for all the hard work they have done in helping me maintain my plants. Your days under the hot summer sun are deeply appreciated.

I would also like to thank Kelli Kochan and Dr. Andrew Hillhouse at the Texas A&M Institute for Genome Sciences and Society (TIGSS) for their technical support

with various projects as well as Dr. Nithya Subramanian at the Texas A&M AgriGenomics Laboratory.

This research would not have been possible without the financial support of Dr. Clare Gill's fellowship grant as well as the support from Dr. Don Jones and Dr. Robert Nichols at Cotton Incorporated. I aim to make returns on their investments in me.

I would also like to thank all my friends and colleagues at the Interdisciplinary Graduate Program in Genetics at Texas A&M University for making this a wonderful experience. I would like to particularly thank Ximena Paez-Colasante and Tamara Ospina-Vega for their diligence in making the Genetics Program a well-running machine and for their dedication to all of its students

I have made life-long friends and have met the most wonderful individuals. So to all of you who have accompanied me throughout this journey, "Thank you!"

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professors David M. Stelly (advisor) and Professor Seth C. Murray of the Department of Soil and Crop Sciences, Professor Clare A. Gill (co-advisor) of the Department of Animal Science, Professor Alan E. Pepper of the Department of Biology, and Professor Mauricio Ulloa of the Department of Plant Stress and Germplasm Development Research at the USDA.

The genotypic data on RIL populations [TM-1 x 3-79 (USDA-ARS, College Station, TX), PS7xNemX (UC Riverside) interspecific and PHY72xSTV474, STV474xPHY72, PHY72xNM67 (USDA-ARS, Lubbock TX) intraspecific] for Chapter II were developed and provided by Dr. Mauricio Ulloa, USDA-ARS Research Geneticist. All other work for the dissertation was completed by the student.

Funding Sources

Graduate study was supported by a dissertation research fellowship from the National Institute of Food and Agriculture, USDA.

This work was also made possible in part by Cotton Incorporated Grant Numbers 13-466TX, 13-694, 17-026, 18-201, and 18-338 as well as by the National Science Foundation Plant Genome Research Program (NSF-PGRP) Grant Numbers IOS1025947, IOS1444552, and IOS1739092.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

Evolutionary History of *Gossypium* Genus

Plants of the cotton genus *Gossypium* are typically woody, perennial shrubs or small trees occurring naturally in tropical and subtropical regions of the world**.** Several species were independently domesticated thousands of years ago in both the New World (the Americas) and Old World (Africa and Asia) [1]. Recent and contemporary cotton production relies on cultivation of 2 allotetraploid ($2n = 4x = 52$) species, both with 26-chromosome AD-genomes, and 2 diploid ($2n = 2x = 26$) species, both with 13-chromosome A-genomes [2]. The vast majority of contemporary production, ca. 95%, relies on the allotetraploid Upland cotton – *G. hirsutum* (L.) – while lesser amounts are produced by *G. barbadense* i.e., Pima and Egyptian cotton, and by the diploid species *G. herbaceum* and *G. arboreum* [2].

The "AD genomes" of extant *Gossypium* allopolyploids descend from a common ancestral genome ($n = 2x = 26$) that arose in the New World approximately 1-2 million years ago (MYA) from interspecific hybridization and the union of two separate genomes, one A-like and one D-like, relative to extant species' genomes [3]. The A subgenomes of extant AD-genome species is related to modern *G. arboreum* (L.) and *G. herbaceum* (L.)*,* both $n = 13$, and originated from Asia, whereas the D subgenome is most closely related to the modern *G. raimondii* (Ulbrich.) ($n = 13$) among the New World D-genome species. While gene content of A and D subgenomes is highly similar, meiotic pairing is considered to be exclusively homologous [4, 5]. Thus, cotton exhibits disomic inheritance and is considered to be a stereotypic allotetraploid. That said, content of the genomes has been greatly complicated by repeated polyploidization prior to formation of the 13-chromosome genome ancestral to modern *Gossypium* genomes [6]. Note that extensive paleo-diploidization has also occurred, and its "footprint" is somewhat variable, as indicated by the presence of the remnant syntenic sequences that allowed detection of the paleo-polyploidization events. Evidence of paleopolyploidization was not only confirmed but extended by comparative DNA sequence analyses across broad dicot taxa, which indicated even higher levels of paleopolyploidy in cotton [7].

The *Gossypium* genus been formally updated to include a new allotetraploid species, *G. stephensii*, from the isolated Wake Atoll island where it was previously labeled as *Gossypium hirsutum* var. *religiosum* (L.) [8]. This brings the total number of allotetraploid species to seven and includes *G. hirsutum* [AD]$_1$, *G. barbadense* (L.) [AD]$_2$, *G. tomentosum* (Nuttall ex Seemann) [AD]$_3$, *G. mustelinum* (Miers ex Watt) [AD]$_4$, *G. darwinii* (Watt) [AD]$_5$, *G. ekmanianum* (Wittm.) [AD]$_6$, and *G. stephensii* (J. Gallagher, C. Grover & Wendel) [AD]$_7$. Including the 7 allotetraploid species, there are 52 diploid species across 8 genomic groups (A-G & K) found across the world [9]. The A-genome diploids include African and Asian cottons which are the origin of spinnable fiber [10] and include the B, E, and F genomes as well. The D-genome diploids are found within Central and Northern America while the remaining C, G, and K genomes are found in Australia.

Importance of Cotton

Cotton is both a major textile fiber crop and an important oilseed crop. The fiber removed from seedcotton by ginning accounts for 30% of textile fiber use world-wide and is produced by over 75 countries, with China, India, and the United States collectively producing two-thirds of all cotton [11]. The United States is the largest exporter of cotton and the seed holds economic value as cattle feed, and as an oilseed – cottonseed oil is used for cooking, soap, and emulsifiers [12]. The cotton produced in the U.S. accounts for more than $21 billion in economic activity [11] and is Texas' leading cash crop. Texas produces 25% of the cotton grown in the U.S. across 6 million acres (23,000 square km) totaling approximately $2.2 billion in cash receipts and $1.6 billion in commodity exports [13].

Need for Genetic Improvement of Upland

Growing threats to the sustainment of agriculture include rising energy costs, availability of highly arable land, increasing temperatures due to climate change, availability of water, and growing pressures from biotic stresses such as insects and fungal infection. The consequences of the aforementioned threats are expected to have a significant impact on Upland cotton production due to its particular sensitivity to increased temperatures, high water requirement, and lack of resistance to various soil pathogens such as certain wilt-causing forms of *Verticillium* and *Fusarium*. Upland cotton must then be prepared to respond to these rapidly changing abiotic and biotic stresses through genetic variation which can be achieved through interspecific germplasm introgression.

Increased temperature has been shown to negatively impact the development of a cotton seedling's root system due to the rapid loss of soil moisture under arid conditions [14] as is common in some production regions, e.g., in Pakistan which depends heavily on cotton production [15]. High temperatures have also been demonstrated to negatively impact canopy growth in both Upland and Pima [16] as well as the reproductive processes leading to fruit development in Upland [17]. In Upland cotton, transgressive segregation of drought-resilient fiber traits have been detected as various QTLs within mapping populations suggesting the genetic potential to maintain fiber quality under drought conditions [18], but advanced recombinant inbred lines were required to detect these QTLs indicating that these traits are segregating infrequently if at all. As such, increasing the rates of recombination or leveraging recombination at specific regions within these lines could expedite the segregation of drought-tolerant traits and genetic improvement of Upland cotton.

An existential threat to Upland cotton is the emergence and spread of the fungal diseases fusarium wilt, *Fusarium oxysporum* f. sp. *vasinfectum* Race 4 (FOV4), and verticillium wilt, *Verticillium dahliae*. Both are fungal pathogens that can result in marginal leaf chlorosis to stunted and damaged root systems in susceptible Upland and Pima lines [19-22]. FOV4 infestation of soils can persist in dormant states for up to 7 years before disease symptoms begin to manifest in susceptible cotton lines making detection of the pathogen difficult. FOV4 is also highly transmissible through contaminated plant debris, seed, and agricultural personal or equipment further exasperating its impact on cotton.

Infection of susceptible Upland lines by verticillium wilt, caused by *V. dahliae*, results in similar disease symptoms as those caused by FOV4. Interestingly, Pima cotton is known to be highly resistant to verticillium wilt as seen in various Pima cottons as well as interspecific Upland and Pima lines [23, 24]. Yet, the development of verticillium resistance Upland lines that maintain their agronomic performance have not been established. The lack of progress in introgressing verticillium resistance while maintaining desired agronomic traits in cotton has been attributed to linkage drag [25].

Origin of Upland Cotton and Its Genetic Improvement

Upland cotton was domesticated approximately at least 5,000 years ago or more[26]. Selection occurred over the centuries for photoperiod-insensitivity and other traits and continued as forms suitable to production in various parts of the USA were sought. At least in more modern times, targets for improvement have included improved fiber characteristics such as increased fiber length and strength, higher lint percentage, and increased uniformity and micronaire. These fiber characteristics have allowed Upland cotton to become a global commercial success, but also have contributed to loss of genetic diversity [27]. Insufficient diversity is thought to limit opportunities for genetic advance for single traits, and to accentuate negative associations between yield and certain fiber attributes, which can arise from inter-related features, effects by common confounding factors and/or linkage drag [28].

One means to overcome loss of diversity among elite forms of a domesticated species and improve subsequent opportunities for genetic gain from selection is to conduct wide-cross germplasm introgression into a cultivated species. The level of difficulty and time required for such efforts varies widely and are generally related to the genetic distance of a germplasm donor to the crop of interest (recipient). Harlan and de Wet (1971) provided simplified classification system of crop "Gene Pools" to facilitate categorization of prospective sources of new germplasm, where "Primary" included sources that were easy to use, e.g., other cultivars and wild forms of the cultivated species, "Secondary" included sources for which transfer is possible but may be difficult and their use requires considerable effort, and "Tertiary" included sources from which transfer was not possible or if so, required radical techniques. For Upland cotton, the "Primary Gene Pool" would include all AD cotton species, as they share a common chromosome number and are readily interbred and backcrossed, while the "Secondary Gene Pool" would include the non-Australian diploids, except, perhaps for the E-genome species [29].

Although all AD species were considered to be in the "Primary Gene Pool" of Upland cotton, the need for genetic diversification has been recognized, pedigree-based analysis of successful cultivars revealed a lack of wide-hybridization and introgression [30]. Of course, this absence could be the result of too little effort, efforts of inadequate duration, and/or use of inadequate or insufficient methodologies. The authors surmised that "Unless progress is made in transferring useful allelic variation from diverse to adapted germplasm without negative agronomic effects, germplasm resources will probably remain underused and the trend towards increased genetic vulnerability will continue". Modern sequencing technologies offer great sensitivity and resolution, so seem likely to provide greater detail and insight into some of the challenges that hamper wide-cross breeding of Upland cotton.

Linkage drag is a barrier to germplasm introgression efforts into Upland cotton from related allotetraploid species. Wide-crossing Upland with wild *G. tomentosum* and *G. mustelinum* largely negates the progress achieved through domestication by introgressing alleles that result in poorer fiber characteristics which are more akin to wild cotton. The classical solution to this problem would be to use subject these interspecific populations to one or more backcrosses followed by inbreeding until alleles responsible for deleterious traits segregate from alleles responsible for agronomically important traits. Segregation would result from meiotic recombination in heterozygous regions. Decreased, or complete lack of segregation of alleles can occur if alleles on the same chromosome are in physical proximity to one another, thus diminishing the probability of a crossover event occurring between the two alleles. This can result in the non-random association of alleles, also known as linkage disequilibrium. The unintentional selection of alleles due to their physical proximity to alleles under selection is known as linkage drag, or genetic hitchhiking, and has been a barrier towards expanding the genetic base of Upland through wild germplasm introgression [25].

Even though Pima cotton is used for wide-cross introgression, the improvement of Upland cotton via breeding with Pima cotton has yet to result in genetically stable progeny that do not break down past the F_2 generation [31, 32]. The loss of fitness in segregating hybrid generations, such as inferior viability and fertility, is known as hybrid breakdown, and along with linkage drag, has made the introgression of germplasm from the related allotetraploid species into Upland cotton difficult. Sequence comparisons between *G. hirsutum* and *G. barbadense* has revealed a total of 10,366 genes, out of approximately 62,274 shared genes (16.7%), with sequence variations between the two species with gains or losses of stop codons and frameshifts [32]. Functional categorization of these genes has also revealed associations with defense response, DNA integration, and cell recognition which has provided insight into the genetic mechanisms responsible for hybrid breakdown between these two species. Other mechanisms responsible for hybrid breakdown which have been proposed include Bateson-Dobzhansky-Muller (BDM) incompatibilities which involves negative epistatic allelic interactions [33], disruption of gene complexes [34], as well as incompatibilities between the nuclear genome and mitochondrial or chloroplast genomes within the cell [35]. The process of hybrid breakdown has also been attributed to speciation in which the early stages of genome divergence begin to manifest themselves as incompatibilities between genomes [35].

Prezygotic barriers resulting in divergent phenotypic characteristics between Upland cotton and wild *G. tomentosum* and *G. mustelinum* may also be responsible for decreased historical hybridization between these genomes. *G. tomentosum* has been isolated in Hawaii, and it was not until the introduction of cotton plantations that contact between *G. tomentosum* and *G. hirsutum* occurred following their evolutionary divergence. *G. hirsutum* and *G. barbadense* flowers are receptive throughout the day and not at night. This is in contrast to *G. tomentosum* flowers which begin to become receptive to pollen in the evening and continue to be receptive into the night [36]. *G. mustelinum* though, has shared a geographical history with *G. barbadense* as well as to introduced *G. hirsutum*, but little introgression of *G. mustelinum* into these two genomes has occurred [37]. *G. mustelinum* also exhibits photoperiod-sensitivity hindering its cultivation in varying latitudes and complicating wide-cross introgression into Upland challenging, at least in some environments.

Need to Understand Linkage and Recombination in *Gossypium* Genus

Sequences that are inherited as a unit due to lack of recombination constitute a linkage block. Logically, these most often involve alleles of genes that are close in physical proximity, but they can be physically distant, too, given a dearth of recombination over the entire segment of separation. Linkage blocks that are separated by one another due to historical recombination are also referred to as haplotype blocks since these contiguous alleles are likely to be maintained and inherited at the haploid stage. Haplotype blocks can also emerge as a result of selective sweeps such as those resulting from domestication. Strong selection for a beneficial allele can result in an increase in its frequency until it becomes fixed, or approaches fixation, within a population [38]. Alleles in close proximity to the beneficial allele are inadvertently selected for as well due to linkage drag. Fixation or near fixation can result in the homogenization of haplotypes due to a decrease in genetic variation which renders recombination ineffective at generating novel allelic combinations at these sites thus resulting in haplotype blocks. Haplotype blocks can also emerge from genetic drift even where recombination crossovers are uniformly distributed, although this is rare and mainly supported by computational models [39]. Understanding the haplotype structure of Upland cotton, as well as the haplotype diversity, is important for understanding the efficacy of recombination and segregation of alleles, as well as for predicting the marker densities required to capture the most genetic variation for genome-wide association studies (GWAS) [40]. Similar ramifications clearly extend to applications of targeted and genome-wide selection.

A common cause of haplotype blocks is diminished recombination along a chromosome. Low rates of recombination can result from low frequencies of crossing over, and/or can be due to the rigid elimination of crossover products by negative selection at gametophytic, endospermic and zygotic stages. Crossovers are non-uniform and occur at elevated rates within the sub-telomeric ends of a chromosome and are largely suppressed in pericentromeric and centromeric regions. The variation in cross over locations along the chromosome is influenced by several genes which are involved with epigenetic deposition, anti-crossover enzymes, and other general regulators [41]. The protein sporulation-deficient 11 (SPO11) is responsible for catalyzing double stranded breaks along a chromosome during prophase I of meiosis. Crossover-rich regions are also associated with hypomethylation in maize [42], *Arabidopsis* [43], and cotton [44] and can be affected by specific mutants such as *met1* and *ddm1*, which result in extensive hypomethylation [45, 46]. Genomic factors such as transposable element insertions are also known to suppresses recombination, because they are prone to becoming methylated by the host to prevent their transposition throughout the genome [47]. In allopolyploids such as *G. hirsutum, Brassica napus,* and polyploid wheat, meiotic crossovers are limited to homologous chromosomes despite the potential for multiple pairing partners. In hexaploid (*Triticum aestivum*) and tetraploid wheat, deletion or mutation of the *Ph1* locus on chromosome 5B results in pairing between homeologous chromosomes [48, 49] and duplication of the *Ph1* locus results in decreased frequencies of multivalents during zygotene [50]. Ultimately, the *Ph1* gene controls the condensation of heterochromatin within several *Triticeae* species and is responsible for conformational changes that allow the recognition of pairing partners during meiosis [51]. Understanding the genetic and genomic features associated with recombination within allotetraploid cotton will facilitate the development of strategies, germplasm and other resources that enable our ability to manipulate recombination, one application of which will be to mitigate the deleterious effects of linkage drag caused by haplotype blocks.

CHAPTER II

HAPLOTYPE STRUCTURE OF *GOSSYPIUM HIRSUTUM*

Introduction

Upland accounts for a majority of the globally produced cotton due to its high yield and environmental adaptability, but its fiber quality is inferior compared to that of the extra-long staple Pima cotton. Pima has greater fiber length, strength, and fineness but its yield potential is hindered by its lower lint percentage and narrow environmental adaptability. Upland cottons' poorer fiber quality is a reflection of its low genetic diversity [52]. Despite their karyotypes, introgression of traits of agronomic importance from Pima into Upland genetic backgrounds has proven to be difficult. Genetic divergence of these two species allows for good vegetative F_1 hybrid vigor, but also leads to reduced fertility and cytological abnormalities in F_2 hybrids and subsequent generations [53]. The decrease in phenotypic performance in subsequent progeny of hybrids is known as hybrid breakdown and may be a result of incompatibility between the genes of different species [31]. This has made the introgression of favorable traits from different species difficult in domesticated cotton.

Decreased recombination, or complete lack thereof, can also result in alleles becoming non-randomly associated with one another, a phenomenon known as linkage disequilibrium (LD). Linkage disequilibrium induced by lack of recombination can cause alleles of neighboring loci to form tightly associated regions that are inherited as a unit which are generally referred to as linkage blocks or haplotypes. These can complicate or even undermine introgression efforts i.e. when a favorable allele within a linkage block is associated with a neighboring deleterious allele. This unfavorable condition is generally referred to as linkage drag. In non-improved types and other species that might be used for wide-cross introgression, there are many alleles ill-suited to agricultural uses and they are distributed widely in the genome[28, 32]. In fact, in most situations of wide-cross introgression, the vast majority of introgressed genes would be expected to be agriculturally deleterious or neutral. Thus, inbreeding of wide hybrid germplasm thus leads to poor agricultural phenotypic performance and hybrid break down. Conversely, marker-defined haplotypes can be associated with well-defined traits and can be used to identify novel recombination within populations facilitating the segregation of favorable from unfavorable alleles [54]. Therefore, understanding the haplotype structures of cotton populations and their associations with desired traits is important for marker-assisted selection in breeding programs.

There are two major statistics used to measure linkage disequilibrium – the square of Pearson's correlation coefficient between pairs of loci (r^2) [55] and Lewontin's normalized coefficient of linkage disequilibrium (*D'*) [56] between pairs of loci. The square of the correlation coefficient (r^2) considers the frequency of the alleles at each locus while Lewontin's normalized D (*D'*) does not and instead is scaled by its theoretical maximum. Ultimately, *D'* is a descriptive statistic that attempts to translate LD to a direct measure of co-inheritance while *r2* is a predictive statistic that utilizes the frequency of alleles to predict the probability of co-inheritance. The former is useful when the sole underlying assumption for the cause of LD is historical recombination, whereas the latter is useful for genome-wide association studies (GWAS) in which the associations between SNPs are useful for estimating the SNP density required to detect an association between a phenotype and a causal SNP (commonly referred to as SNP tagging). Although both statistics can be used to estimate haplotype structures within a genome, Lewontin's *D'* has been used as a proxy for historical recombination between SNPs and has been adapted into three main haplotype partitioning models which in turn have been modified to handle larger datasets and increases in computational capability [57].

The simplest method partitions SNPs into haplotypes by calculating *D'* between two neighboring loci and extends the comparison until a pair of loci have a *D'* less than a pre-specified threshold and is referred to as a solid spine (SS) of LD [58]. The second method infers that historical recombination has occurred between a pair of loci if all four haplotypes, or gametes, are observed with at least 1% frequency and is referred to as the four gamete test [59]. Lastly, the third method relies on the confidence bound of *D'* rather than its actual estimate where the upper 95% confidence bound has a *D'* greater than 98% and the lower confidence bound of *D'* greater than 80% [60]. This method is referred to as confidence intervals (CI) and exhibits increased robustness against erroneous SNPs generated by genome assembly errors, gene conversions, recurrent mutations, and genotyping. Ultimately, the CI method for haplotype block portioning was used in this study.

It has traditionally been considered that the process of domestication, whether a crop or animal, results in the reduction of its genetic diversity and leads to domestication syndrome – the collection of phenotypic traits associated with the changes involved from transitioning from a wild progenitor genotype to a domesticated genotype [61]. This rationale stems from the classical interpretation of genetic bottlenecks, which proposes that the selection of a few individuals from a larger population, e.g., for cultivation, results in the reduction of genetic diversity compared to the original population followed by genetic stagnation and cessation of gene flow. Efforts to incorporate exotic germplasm into breeding programs have increased as a response to the reduction of genetic diversity in crops due to these domestication-induced bottlenecks [62]. This notion has largely been challenged as the sequencing of the archeological genomes of barley, maize, and sorghum have revealed that the expected drop in genetic diversity did not follow their domestication [63]. On the contrary, a diversity analysis of cultivated and non-cultivated germplasm of *G. hirsutum* revealed that cultivated germplasm exhibits a decrease in genetic diversity compared to wild germplasm [64]. This study aims to determine if there is a decrease in genetic diversity, as measured by haplotype diversity, in areas associated with domestication as inferred by the presence of selective sweeps between improved and wild cotton.

Materials and Methods

Plant Material and Genotyping

Two interspecific mapping populations consisting of 59 *G. hirsutum* (TM-1) x *G. mustelinum* BC_1F_1 and 85 *G. hirsutum* (TM-1) x *G. tomentosum* BC_1F_1 where grown over the summer of 2016 in College Station, Tx. Young leaf tissue was collected from these populations and DNA was isolated using a Machery-Nagel Plant Nucleo-spin kit (Pennsylvania) according to the manufacturer's instructions. Isolated DNA was quantified using PicoGreen® and then diluted to 50ng/µl. Genotyping was conducted using the CottonSNP63K Array at Texas A&M University [65] and SNPs were called using Illumina's® GenomeStudio (v2.0) software resulting in 63,058 SNPs functional and non-functional SNPs per population.

Genotyping data from 3 interspecific mapping populations consisting of 195 TM-1 x 3-79 F_2s which was an expansion of a previous study [65], 76 Pima S7 (PS7) x New Mexico (NemX) RILs (F_8), 69 TM-1 x 3-79 RILs (F_8), and 4 intraspecific mapping populations that collectively involved three parents: Phytogen 72 (PHY72), Stoneville 474 (STV474), and NM67. The 4 intraspecific populations consisted of 93 PHY72 x STV474 F_2s, 132 PHY72 x STV474 RILs (F_8), 104 STV474 x PHY72 RILs (F_8), and 131 PHY72 x NemX67 RILs which were provided by Dr. Mauricio Ulloa from the USDA-ARS (Lubbock, TX). The genotypic data on RIL populations [TM-1 x 3-79 (USDA-ARS, College Station, TX), PS7xNemX (UC Riverside) interspecific and PHY72xSTV474, STV474xPHY72, PHY72xNM67 (USDA-ARS, Lubbock TX) intraspecific] were developed and will be published in collaboration with Dr. Mauricio Ulloa, USDA-ARS Research Geneticist; third parties seeking access to these data before publishing and publicly available should contact Dr. Ulloa.

Development and Mapping of Intraspecific and Interspecific Linkage Maps

Genotyping data were processed for quality control by removing SNPs that had more than 10 percent missing data or had a minor allele frequency (MAF) less than 5 percent. Individuals with missing data greater than 10 percent were also removed. Genotype data were converted to ABH format using a panel of parental genotypes [65] where "A" represents an allele descendant of parent 1, "B" an allele descendent of parent 2, and "H" represents a heterozygous genotype.

Processed genotype data were imported into JoinMap 4.1 [66] and linkage groups were coalesced using the independence logarithm of the odds (LOD) function within the grouping tree tool. The grouping processes was reiterated with increasing LOD thresholds until 26 coherent linkage groups were established. SNPs that did were not integrated into a linkage group were removed. The maximum likelihood algorithm with a maximum recombination frequency of 50% and the Kosambi function were used to order the SNPs within each of the linkage groups. SNPs with a nearest neighbor stress (NNS) greater than 5 centiMorgans (cM) were removed and reordering of markers was conducted until markers exhibited a NNS less than 5. Linkage disequilibrium of the final linkage maps were visualized using the R package ASMap [67]. Array ID sequences from SNPs retained in the final linkage maps were aligned to the JGI G. *hirsutum* v2.0 sequence assembly using BLASTn (v2.7.1+) [68] with a minimum e-value cutoff of 1e-10 and with the dust and soft masking parameters disabled. Array IDs that aligned to both a homeolog relative to the linkage map and to the linkage map homolog were ultimately mapped to the homolog i.e. the linkage map was the deciding factor in ambiguously aligned markers. Array IDs that mapped to the same location were filtered based on BIT score.

Genotyping of Improved and Wild Cotton Accessions

Genotyping data from a *G. hirsutum* diversity panel consisting of 257 improved accessions and 71 wild accessions was acquired from a previously published study [64] generated using the CottonSNP63K Array at Texas A&M University. Genotype data were filtered by removing SNPs with a MAF less than 5% or that had missing data greater than 10%. Homeologous SNPs (homeo-SNPs) that occur due to intragenomic sequence identity were also removed and only markers that were categorized as functionally polymorphic were retained [65]. Array IDs were similarly aligned to the JGI *G. hirsutum* v2.1 sequence assembly using BLASTn. Array IDs that mapped to multiple chromosomes were corrected for using the previously generated linkage maps if applicable. The final genotype data set was imputed and phased using Beagle (v4.1) [69] using the default parameters.

Haplotype Structure and Frequency Estimations

Linkage disequilibrium (LD) analysis using Lewontin's normalized *D'* value was conducted on the final genotype dataset. Haplotype block partitioning was conducted with PLINK (v1.90) [70] using confidence intervals (CI) which classifies pairs of markers into one of three LD categories [60]. Default CI parameters were used with the upper and lower 95% confidence bounds set to 0.98 and 0.70, respectively, and the upper confidence bound of normalized *D'* was set to .90 as evidence for historical recombination. No maximum block length was set, thereby allowing for chromosome-wide haplotype block partitioning. A non-parametric Wilcoxon rank-sum (Mann-Whitney) unpaired test was conducted using the software R between the improved and wild haplotype block lengths following a Shapiro-Wilk test for normality. Haplotype block structure and test results were visualized using the R package ggplot2 while visualization of the haplotype structure of chromosome A08 was conducted using HaploView (v4.2).

The frequencies of haplotypes comprising individual haplotype blocks were estimated using the command line version HaploView (v4.2) [58] on Linux. Haplotypes were estimated using the same identical CI parameters as were used in delimiting the haplotype blocks with PLINK; an in-house script was used to pair the haplotype frequencies with their respective haplotype block. A non-parametric Wilcoxon rank-sum (Mann-Whitney) unpaired test was conducted using the software R between the number of haplotypes per block between the improved and wild populations following a Shapiro-Wilk test for normality. Haplotype estimates and test results were visualized using the R package ggplot2.

Validation of Chromosome A08 Haplotype Structure

Validation of the haplotype structure of improved *G. hirsutum* was performed by conducting paired-end sequencing on 24 improved *G. hirsutum* accessions. Paired-end sequencing was performed using a target-capture based approach that utilized biotinylated 120-mer probes that complement a segment of 80,000 unique sequences designed to target approximately 175,000 Sanger fasta sequences from bacterial artificial chromosomae-end sequences [71]. Generated FASTQ sequence files were analyzed for quality control using the Linux-based software FastQC [72]. The first 12 base pairs (bp) of each fastq file were trimmed and base pairs with a phred score less than 28 were removed using the software fastp (v0.20.0) [73] in paired-end mode to ensure proper paring of reads were maintained. Only reads with a minimum length of 80 bp, out of 100 bp reads, were kept using the same software. Processed FASTQ files were aligned to the JGI *G. hirsutum* v2.1 sequence assembly using the software BWA (v0.7.17-r1188) [74] using the BWA-mem algorithm in paired-end mode. Aligned sequence reads were piped into samtools (v1.7) [75] as binary alignment (BAM) files where they were merged and sorted. BAM files with missing mate pairs were corrected and duplicate reads were removed using samtools. Variant calling was performed using samtools mpileup and piped into bcftools [76] for variant filtering. An in-house script was used to filter out potential homeo-SNPs by removing SNPs that were heterozygous between the TM-1 accession and the reference allele in order to remove heterozygous calls between the identical lines. Individual variants that had a depth of coverage of less than 3 reads were removed using the software vcftools (v0.1.15) [77]. Data were converted into MAP/PED format for PLINK processing. Markers that had a MAF of less than 5% or that had more than 5% missing data were removed using PLINK. Individuals that had more than 10% missing data were also removed. Filtered samples were then used to estimate the haplotype structure of chromosome A08 using CI with identical parameters as previously stated and the haplotype structure of chromosome A08 was visualized using HaploView (v4.2).

Recombination Rate Estimations

The genetic (cM) and physical (bp) positions of the nine previously generated linkage maps were used to estimate the recombination rates of these populations via regression modeling where the recombination rate was estimated as a function of genetic distance over physical distance (cM/Mb). Markers that were non-colinear between the genetic maps and the JGI *G. hirsutum* v2.1 sequence assembly were masked by removing the markers with the greatest pairwise residual between both genetic and physical vectors until both vectors had a pairwise residual of zero. Data with masked markers were imported into the R package MareyMap [78, 79]. Three interpolation methods were modeled to identify the method with the least amount of recombination rate variance: 1) a non-linear locally estimated scatterplot smoothing (LOESS) method with a span of 0.075 and 1st degree polynomial curve where the size of the regression window is a percentage of the total number of markers allowing for variations in marker density; 2) a non-linear cubic spline with automatic estimation of spar (equivalent of span in LOESS) and degrees of freedom with generalized cross-validation; and 3) a linear non-overlapping 1 mega-base (Mb) sliding window requiring a minimum of 4 markers for regression estimation. Final regression data were visualized using the R package ggplot2 [80]. Negative interpolations within the non-linear models were set to a value of zero and sliding windows within the linear regressions that had less than 4 SNPs were set to “NA” as to not provide a pseudo recombination rate of zero.

Selective Sweep F_{ST} Analysis

Unprocessed genotype data of the 257 improved and 71 wild *G. hirsutum* accessions were pooled together and markers that had a MAF of less than 5% or had more than 10% missing data were removed using PLINK. Individuals that had more than 10% missing data were also removed. Markers that had not been previously mapped were aligned to the JGI *G. hirsutum* v2.1 sequence assembly using the previously mentioned methods. Processed PLINK map/ped files were converted to PLINK binary files and used to calculate the number of subpopulation clusters within the pooled cotton panel using the software Admixture (v1.3) [81]. A cross-validation (CV) procedure was used to perform 5-fold CV estimations using the default parameters and was reiterated for 25 individual *K*-mean values consisting of 1 through 25. The *K*-mean value with the least CV error was estimated to be the number of subpopulation clusters within the pooled cotton panel.

The processed binary PLINK files were then used to estimate selective sweeps between the improved and wild cotton accession populations by comparing allele frequencies between populations using the software hapFLK (v1.4) [82]. Selective sweeps where estimated while considering the hierarchical structure of the populations (FLK statistic) [83] and by regrouping individual markers into local haplotype clusters by using the best K-mean value estimated by Admixture. The hapFLK statistics were estimated from the FLK results by using the provided python script. The false discovery rate (FDR) for the resulting p-values were corrected for using the Benjamini-Hochberg procedure using the software R. The resulting p-values and FDR corrected q-values of the identified hapFLK selective sweeps where graphically displayed as Manhattan plots using the R package CMplot (v3.6.0) [84]. The haplotype cluster frequencies for chromosome D08 (c25) were estimated using the provided R script which was modified for visual scaling.

Results

Number of Mapped SNPs via Linkage Mapping and Sequence Alignment

Linkage mapping of 5 interspecific and 4 intraspecific cotton populations resulted in the genetic mapping of 29,149 unique CottonSNP63K Array markers out of the available 63,058 markers (**Table 2.1**). All linkage maps with the exception of STV474 x PHY72 RIL population generated 26 coherent linkage groups each representing a chromosome with no major or minor structural arrangements as demonstrated by a lack of low LOD scores between marker pairs as well as a lack of low recombination fractions between marker pairs not adjacent to one another (**Figure 2.1**). Linkage groups 18, 19, 25, 26 within the STV474 x PHY72 RIL population were split into two smaller linkage groups that were joined based on positional data from sequence assembly alignment (**Figure 2.2**). The Alignment of the diversity panel marker array IDs to the JGI *G. hirsutum* v2.1 sequence assembly resulted in the alignment 24,260 unique markers of which 19,856 markers belonged to the improved cotton panel and 22,381 markers belonged to the wild cotton panel. (**Table 2.1**). A total of 19,709 unique marker arrays IDs that were successfully mapped in the linkage mapping populations and were also mapped within the diversity panel (**Figure 2.3**). A total of 2,510 SNPs (12.7%) and 2,6252 (11.9%) SNPs in the improved and wild cotton panels, respectively were corrected for using the mapping locations provided by the linkage mapping populations (**Figure 2.4**). Only 185 (0.93%) and 168 (0.75%) of the markers within the improved and wild cotton panels could not be corrected for relative to the linkage mapping positions.

Table 2.1 Summary results of the 9 mapping populations and 2 diversity panels.

Population Type	Population	Number of Individuals	Number of SNPs
Interspecific	*G. hirsutum* (TM-1) x *G. barbadense* (3-79) F_2	195	18,659
	G. hirsutum (TM-1) x *G. tomentosum* BC_1F_1	85	14,622
	G. hirsutum (TM-1) x *G. mustelinum* BC_1F_1	59	15,825
	G. hirsutum (TM-1) x *G. barbadense* (3-79) RILs*	90	15,846
	G. barbadense (PS-7) x *G. hirsutum* (NemX) RILs*	115	15,761
Intraspecific	*G. hirsutum* (PHY72) x *G. hirsutum* (STV474) F_2	93	7,171
	G. hirsutum (PHY72) x *G. hirsutum* (STV474) RILs*	132	7,059
	G. hirsutum (STV474) x *G. hirsutum* (PHY72) RILs*	104	6,319
	G. hirsutum (PHY72) x *G. hirsutum* (NM67) RILs*	131	6,198
Diversity Panel	Improved	257	19,856
	Wild	71	22,381

*Third parties seeking access to these data before publishing and publicly available should contact Dr. Mauricio Ulloa.

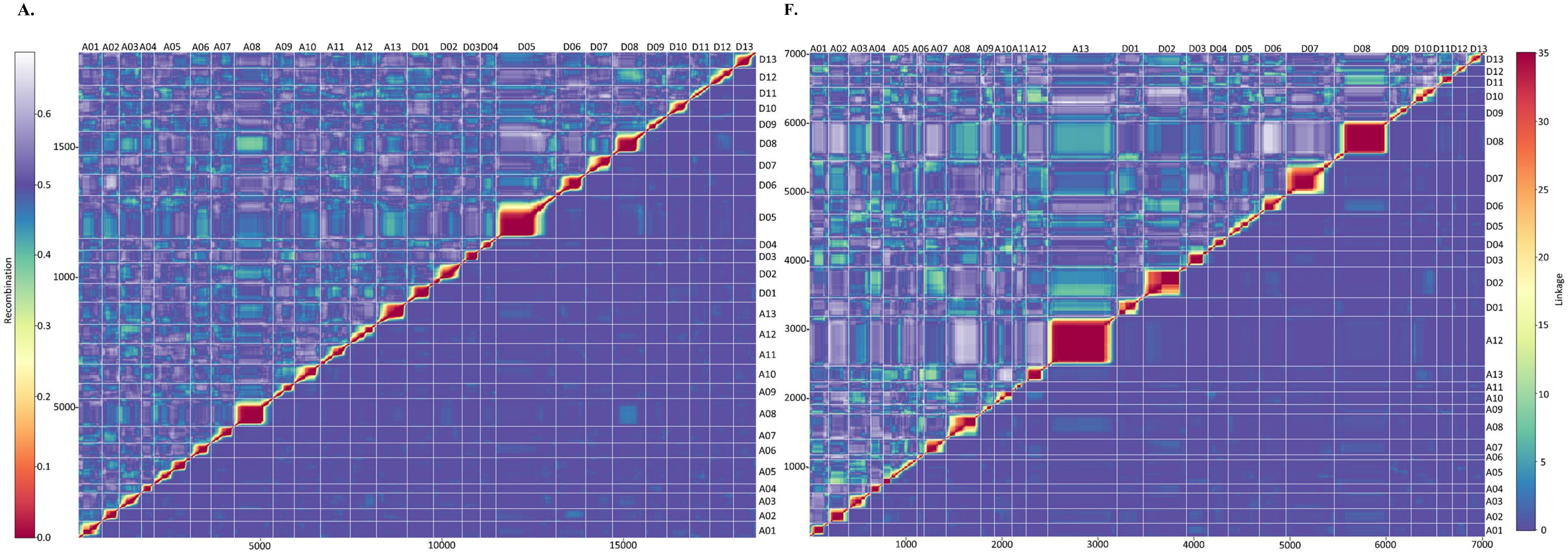

Figure 2.1 Linkage disequilibrium heat maps of mapping populations. Values left of the diagonal represent LOD scores while values right of the diagonal represent recombination fractions. Interspecific populations **A)** GhxGb F_2, **B)** GhxGt BC_1F_1, **C)** GhxGm BC_1F_1, **D)** GhxGb RIL, **E)** PS7xNemX RIL, and intraspecific populations **F)**, PHY72xSTV474 F_2, **G)** PHY72xSTV474 RIL, **H)** STV474xPHY72 RIL, and **I)** PHY72xNM67 RIL are shown. Cold color represents a low value while a hot color represents a high value.

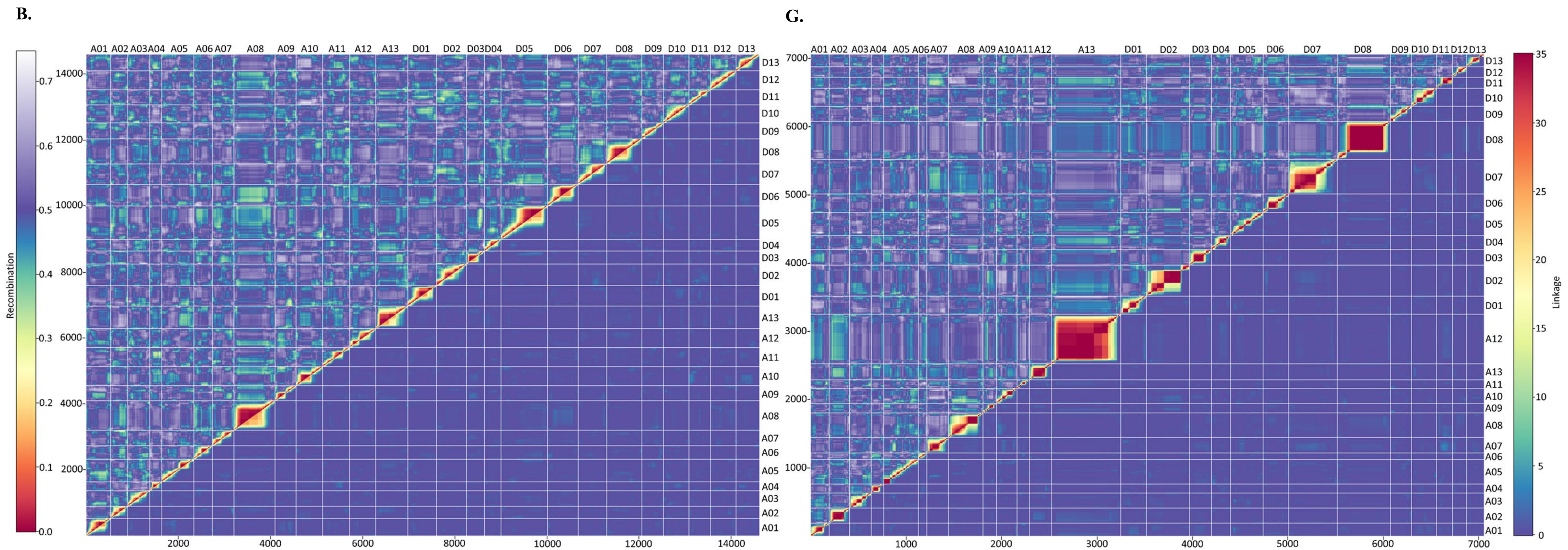

Figure 2.1 (continued)

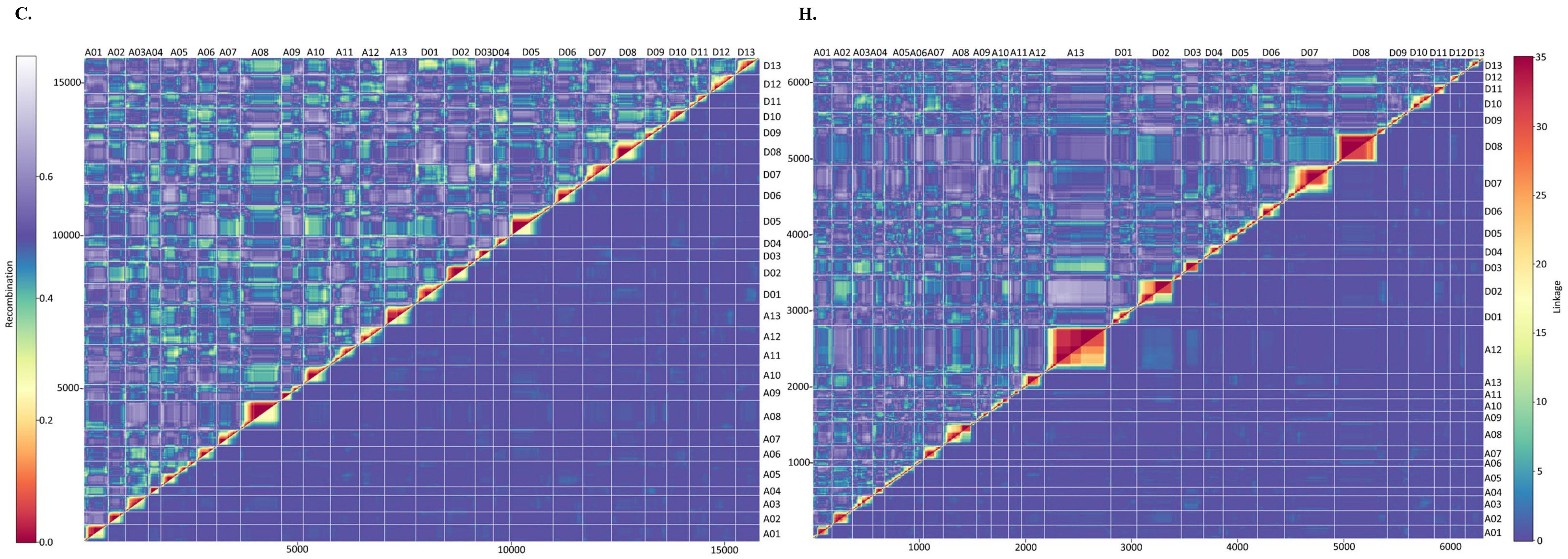

Figure 2.1 (continued)

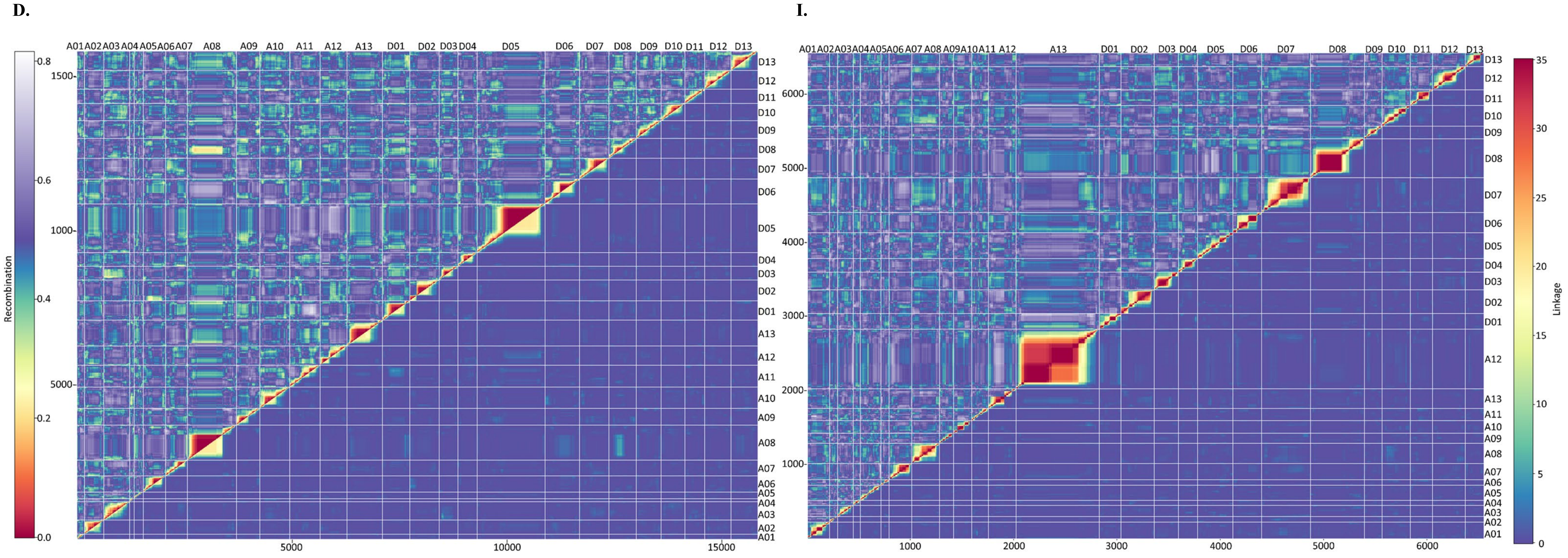

Figure 2.1 (continued)

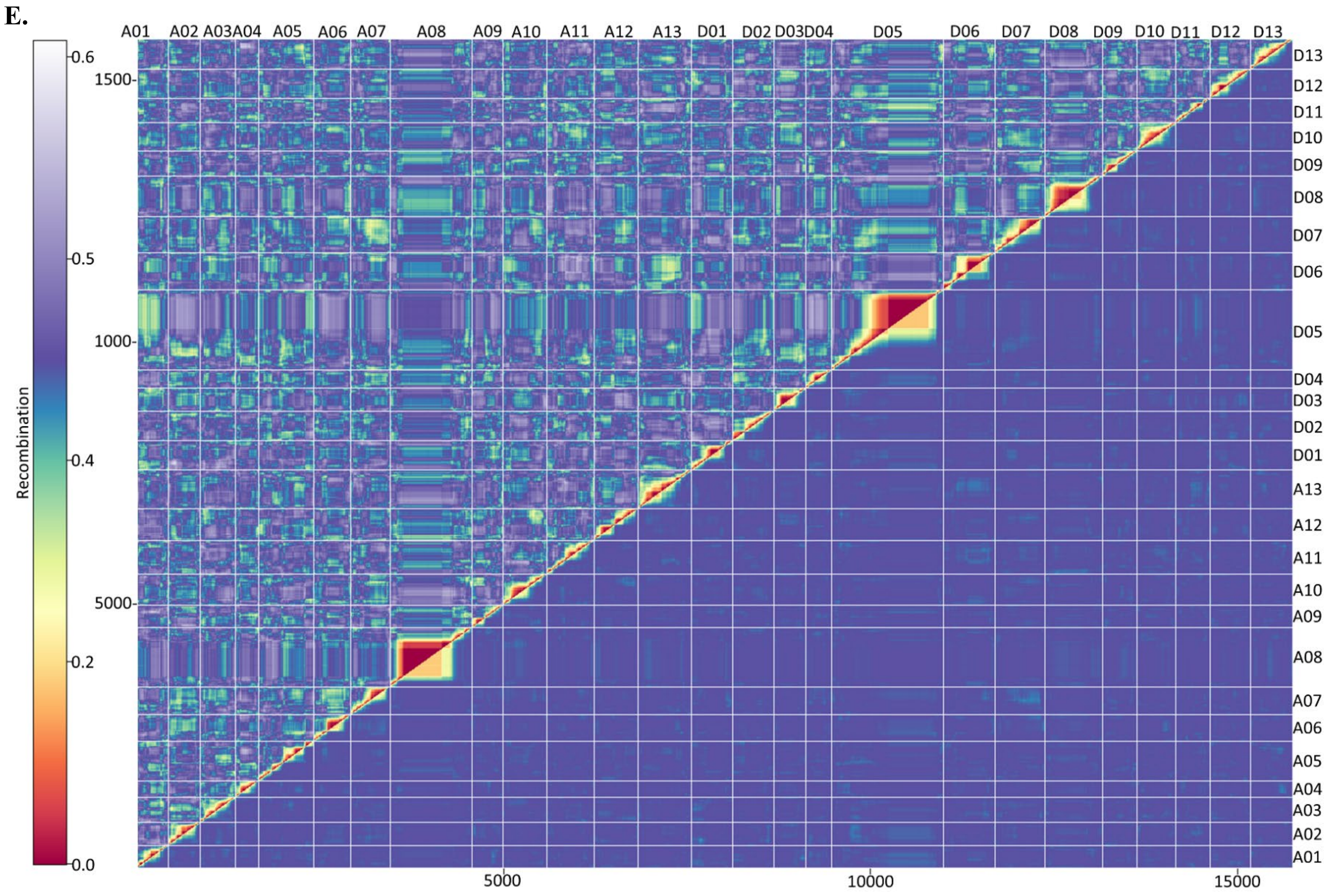

Figure 2.1 (continued)

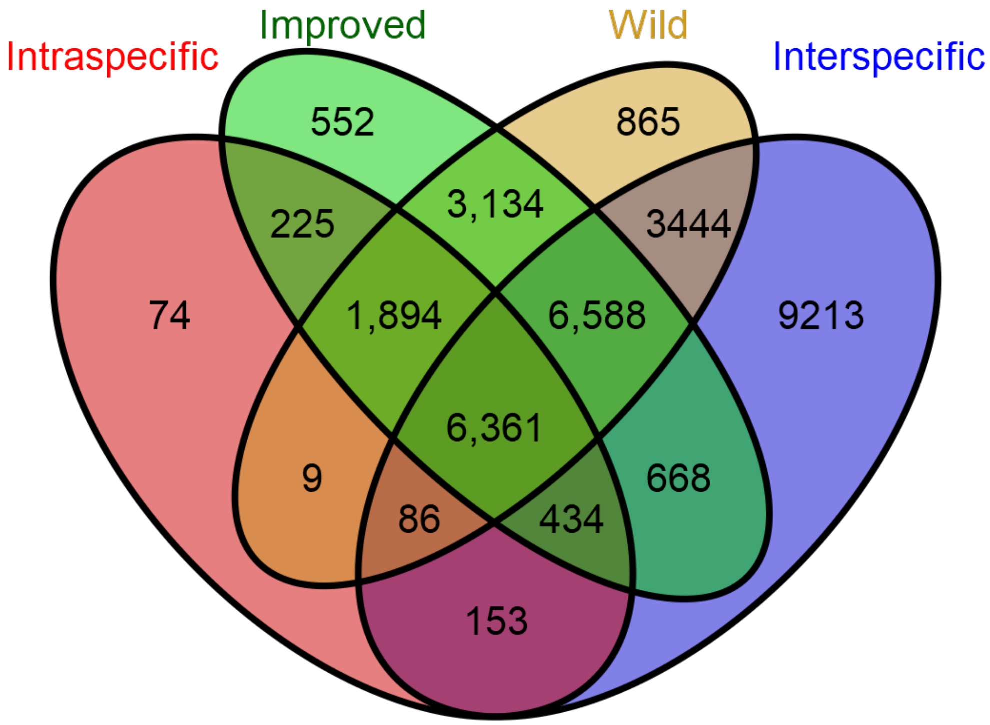

Figure 2.2 Numbers of SNP markers shared between and among the five types of mapping populations and mapped to the JGI *G. hirsutum* v2.1 sequence assembly.

A.

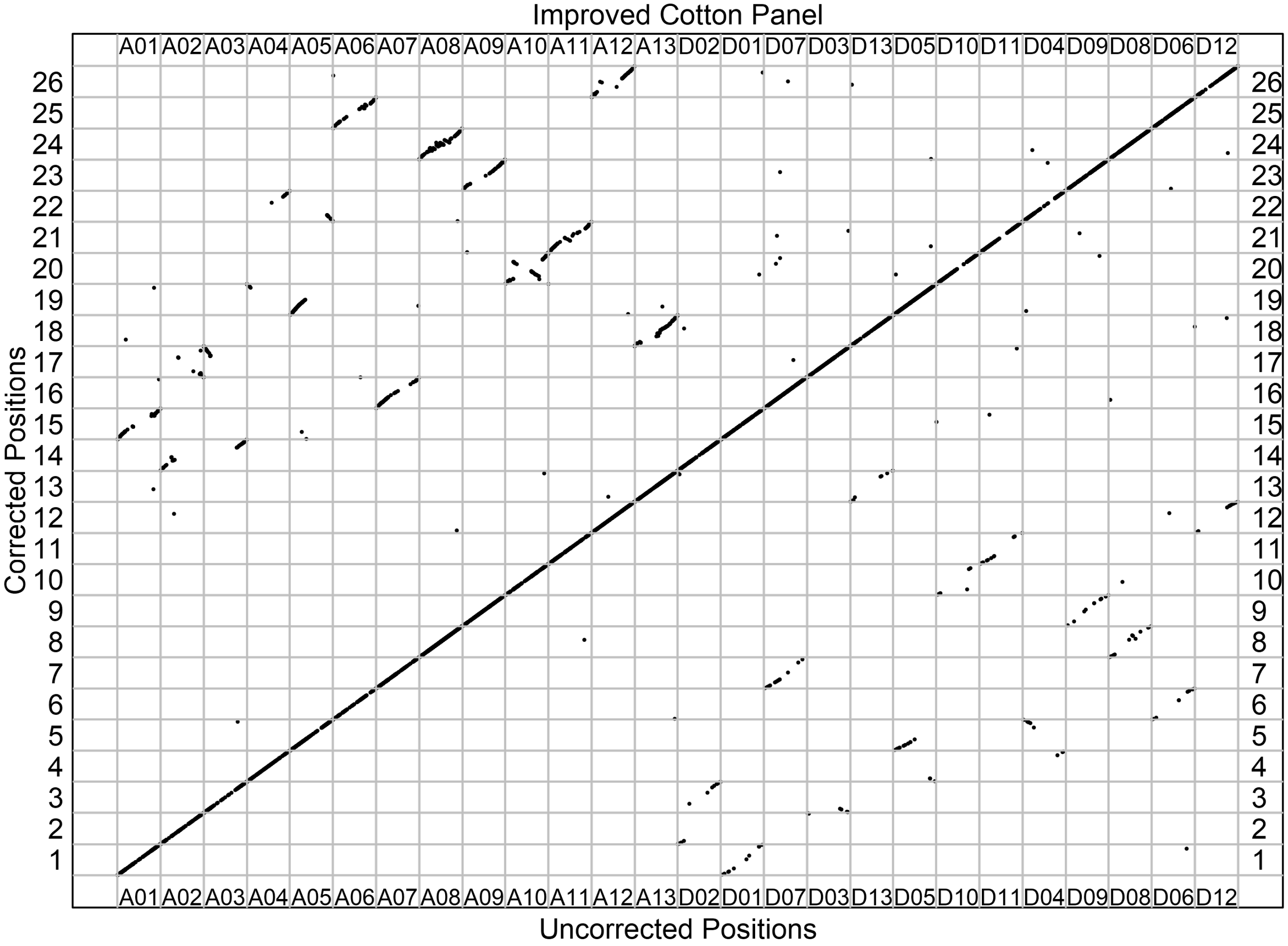

Figure 2.3 Genome wide marker alignment results before and after correction using the mapping populations. A) alignment of the improved cotton panel and **B)** wild cotton panel to the JGI *G. hirsutum* v2.1 sequence assembly. All dots off of the diagonal line represent a corrected SNP position. Original, uncorrected alignment position are shown on the x-axis while the corrected alignment positions are shown on the y-axis.

B.

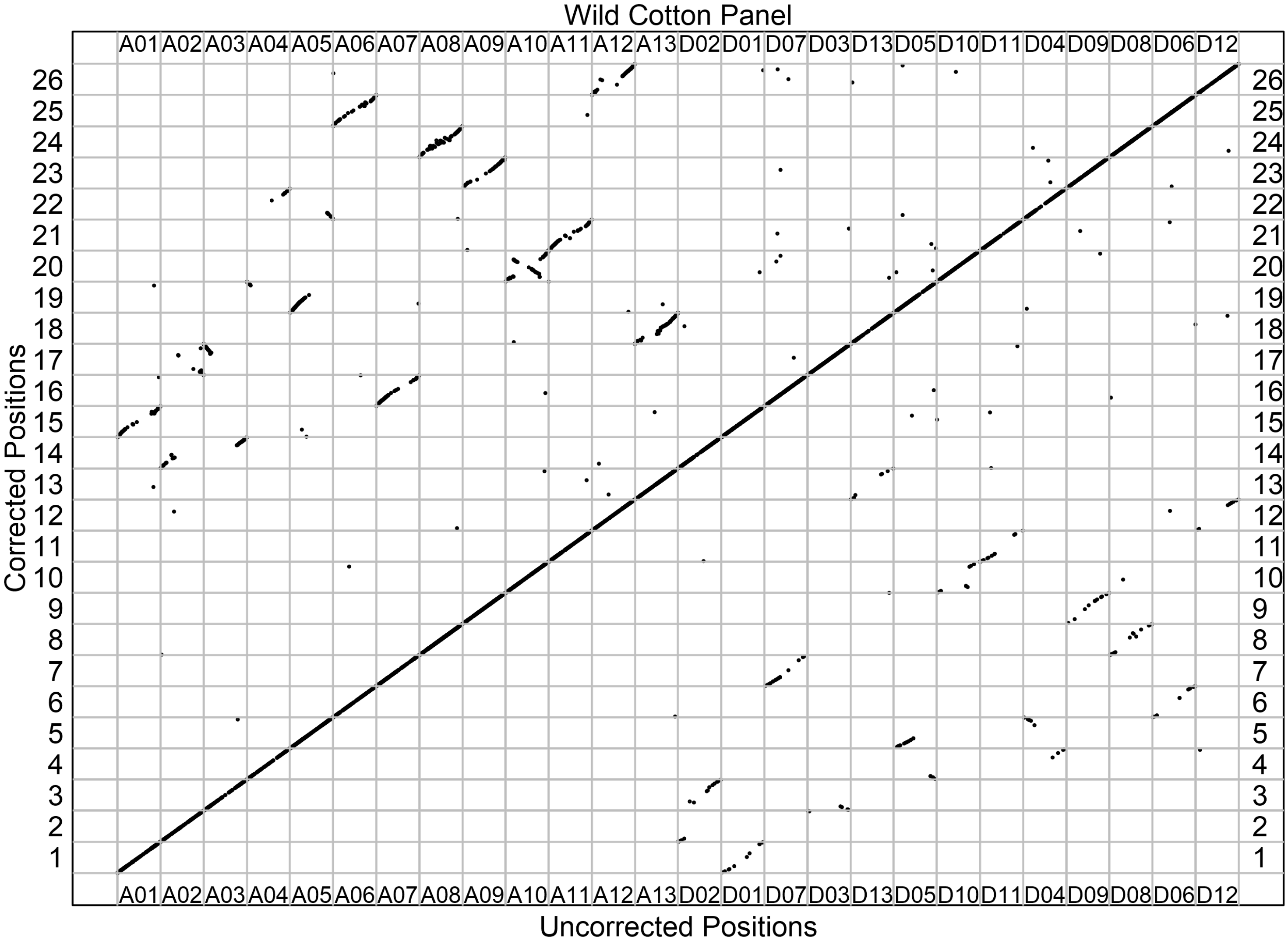

Figure 2.3 (continued)

A.

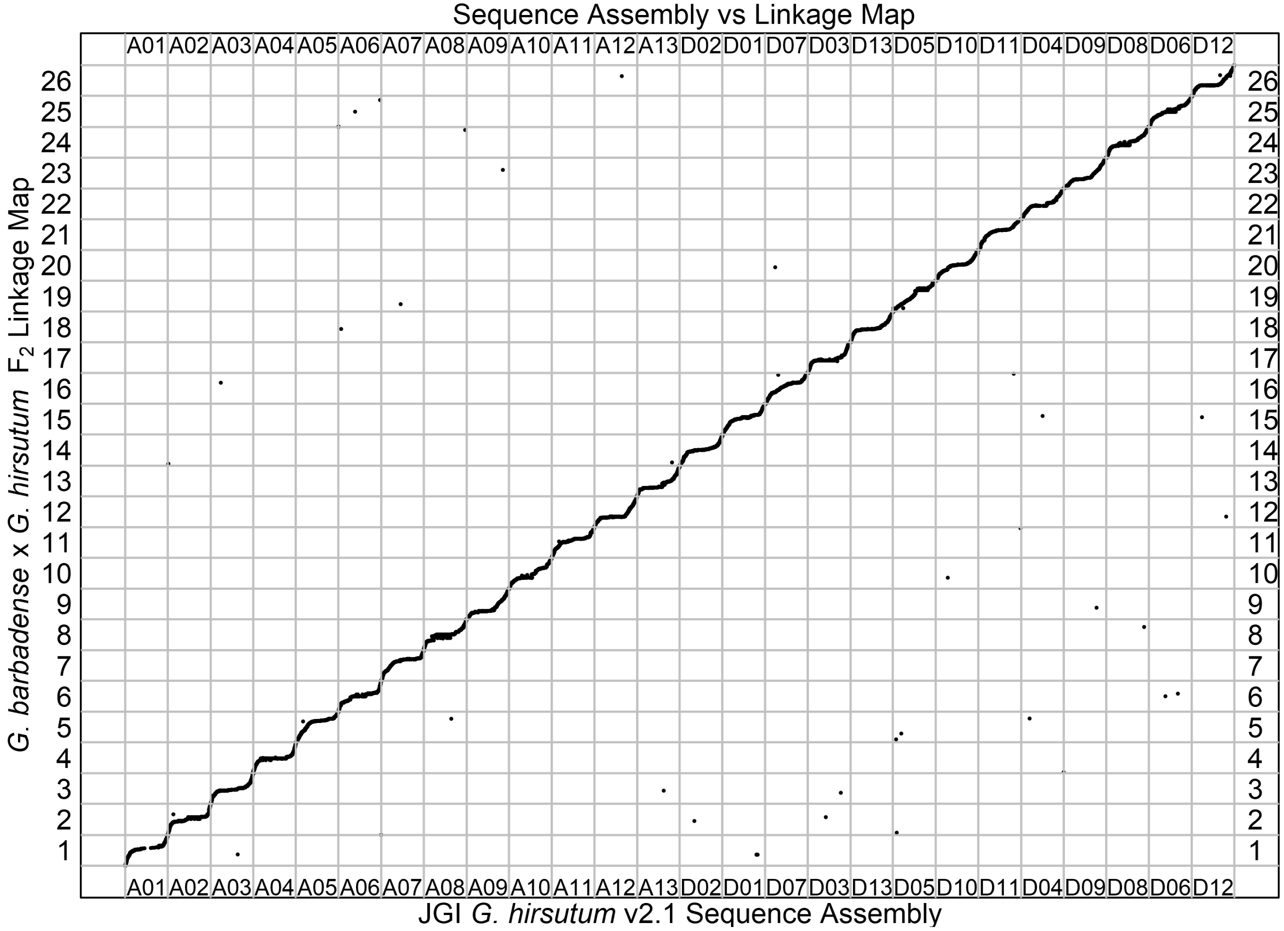

Figure 2.4 Alignment results of the 9 genetic mapping populations to the JGI *G. hirsutum* v2.1 sequence assembly. Interspecific populations **A)** GhxGb F_2, **B)** GhxGt BC_1F_1, **C)** GhxGm BC_1F_1, **D)** GhxGb RIL, **E)** PS7xNemX RIL, and intraspecific populations **F)**, PHY72xSTV474 F_2, **G)** PHY72xSTV474 RIL, **H)** STV474xPHY72 RIL, and **I)** PHY72xNM67 are shown. Sequence assembly position is represented by the x-axis while the linkage mapping position is represented by the y-axis.

B.

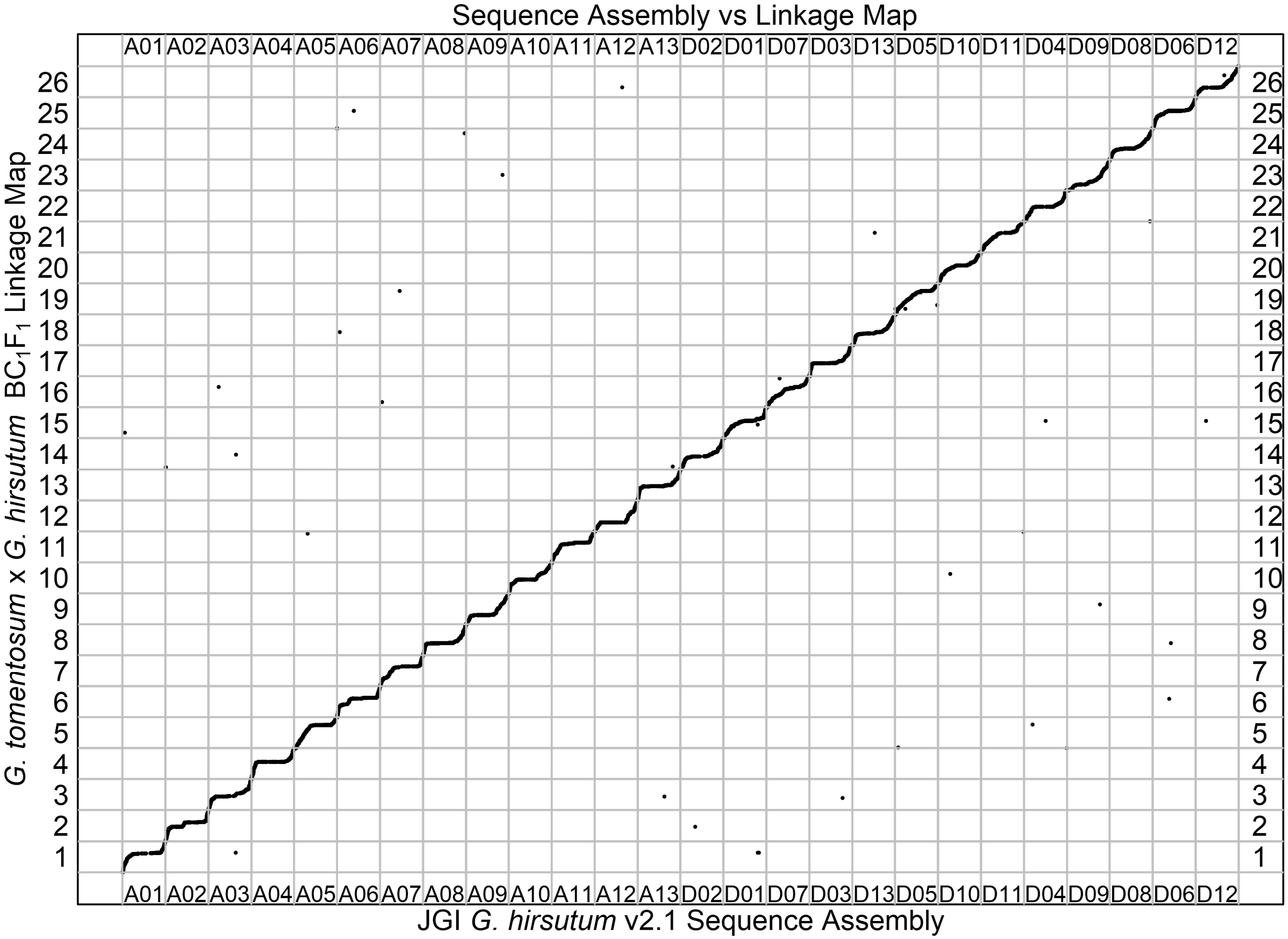

Figure 2.4 (continued)

C.

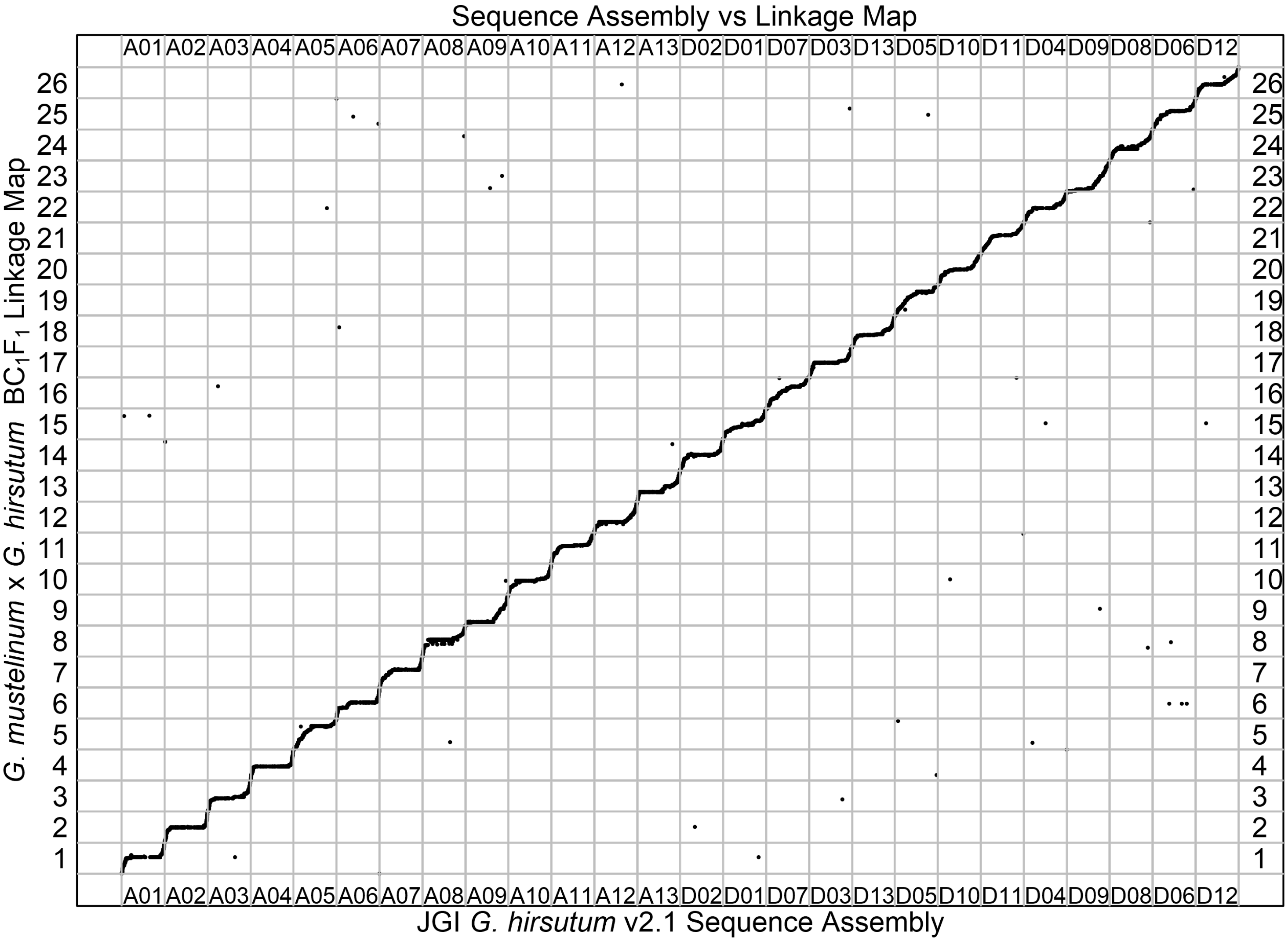

Figure 2.4 (continued)

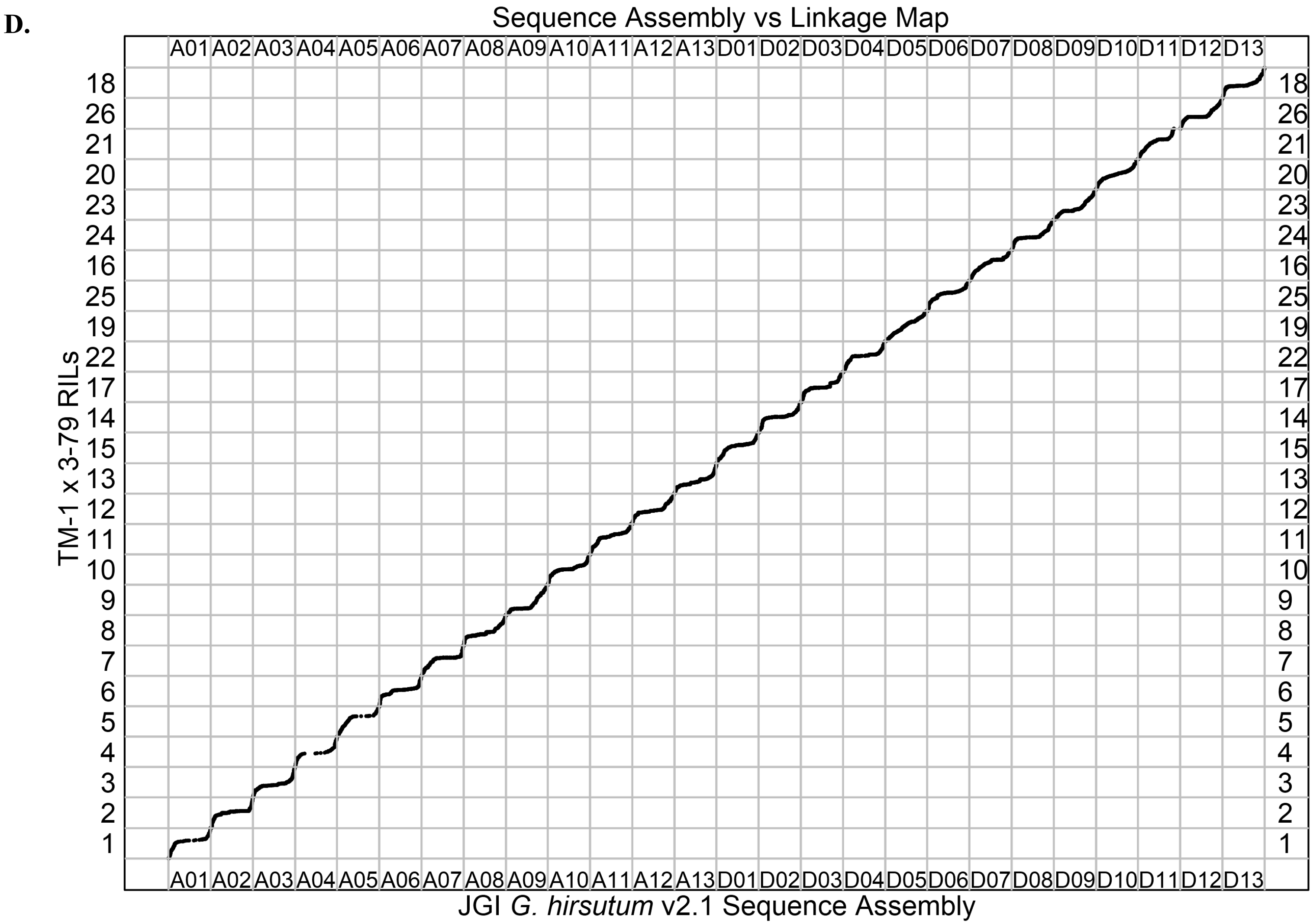

Figure 2.4 (continued)

E.

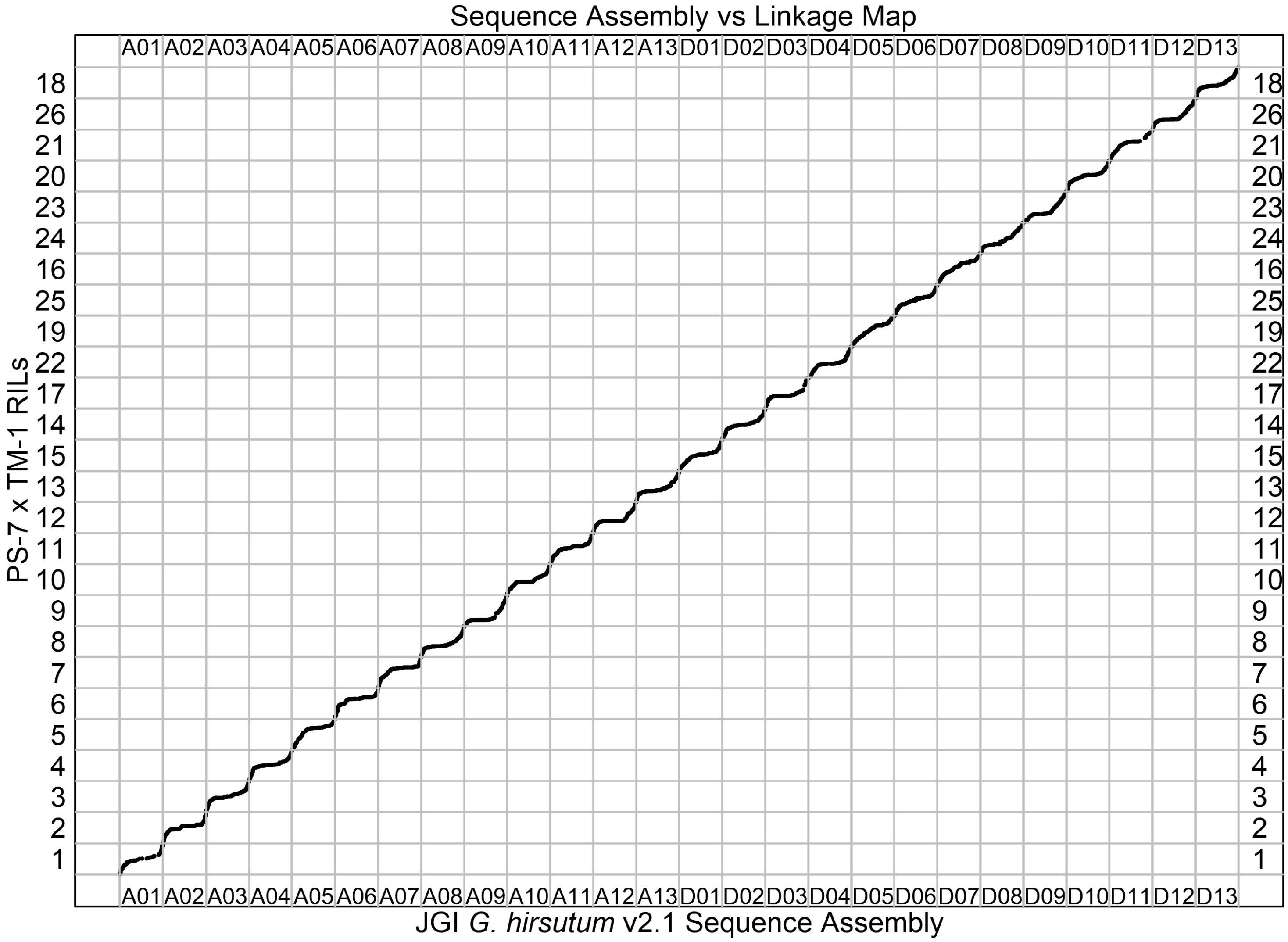

Figure 2.4 (continued)

F.

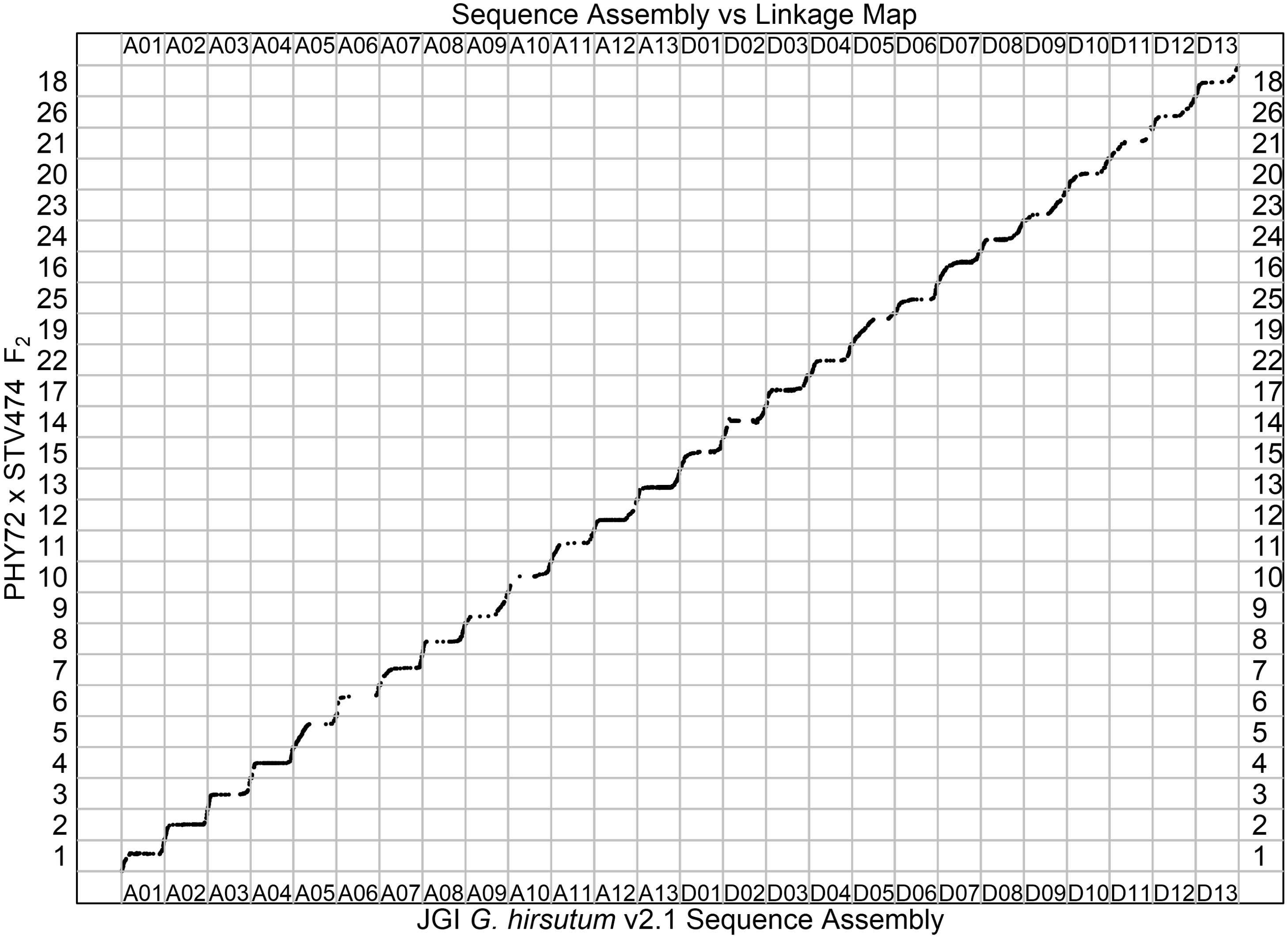

Figure 2.4 (continued)

G.

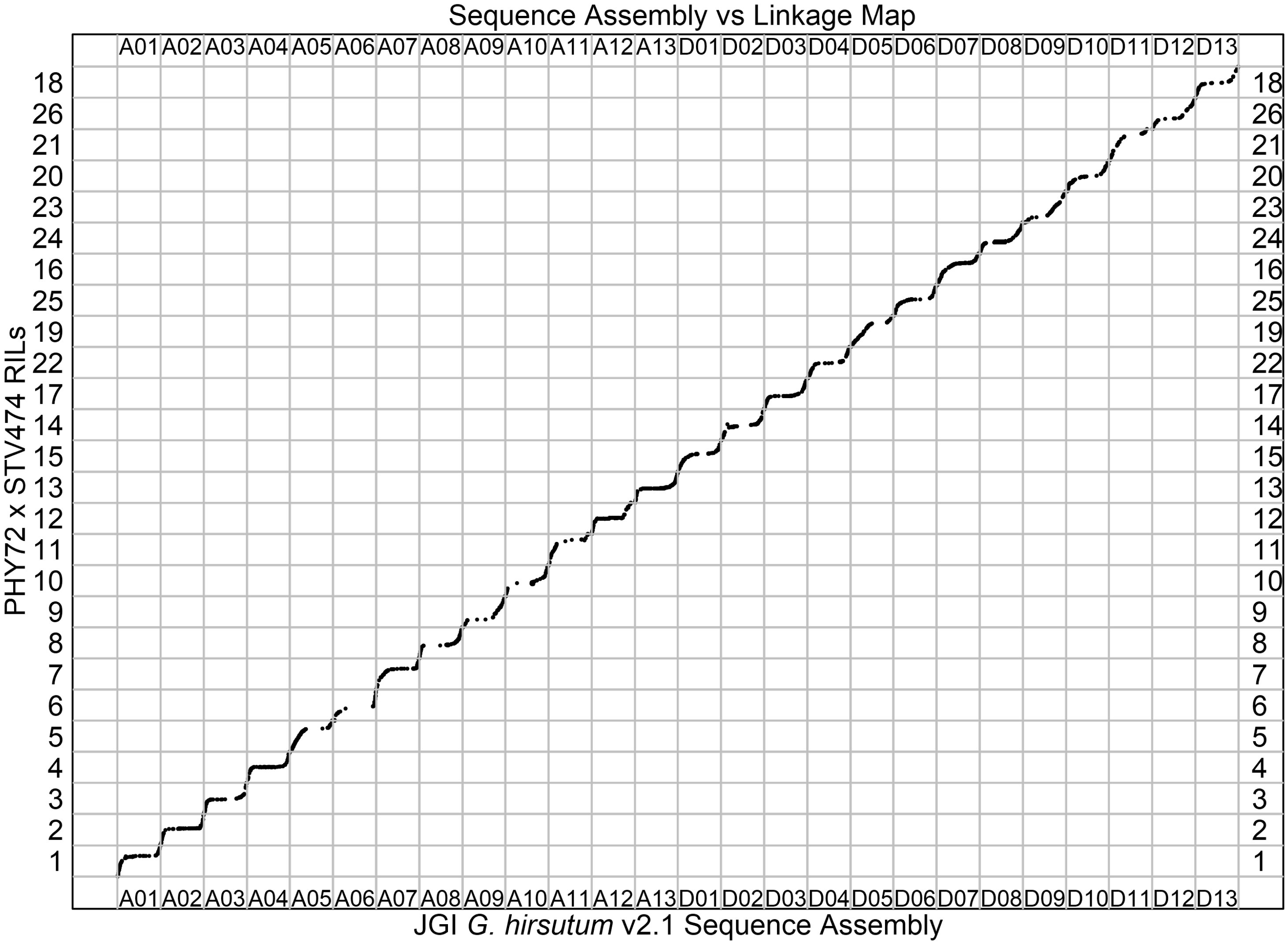

Figure 2.4 (continued)

H.

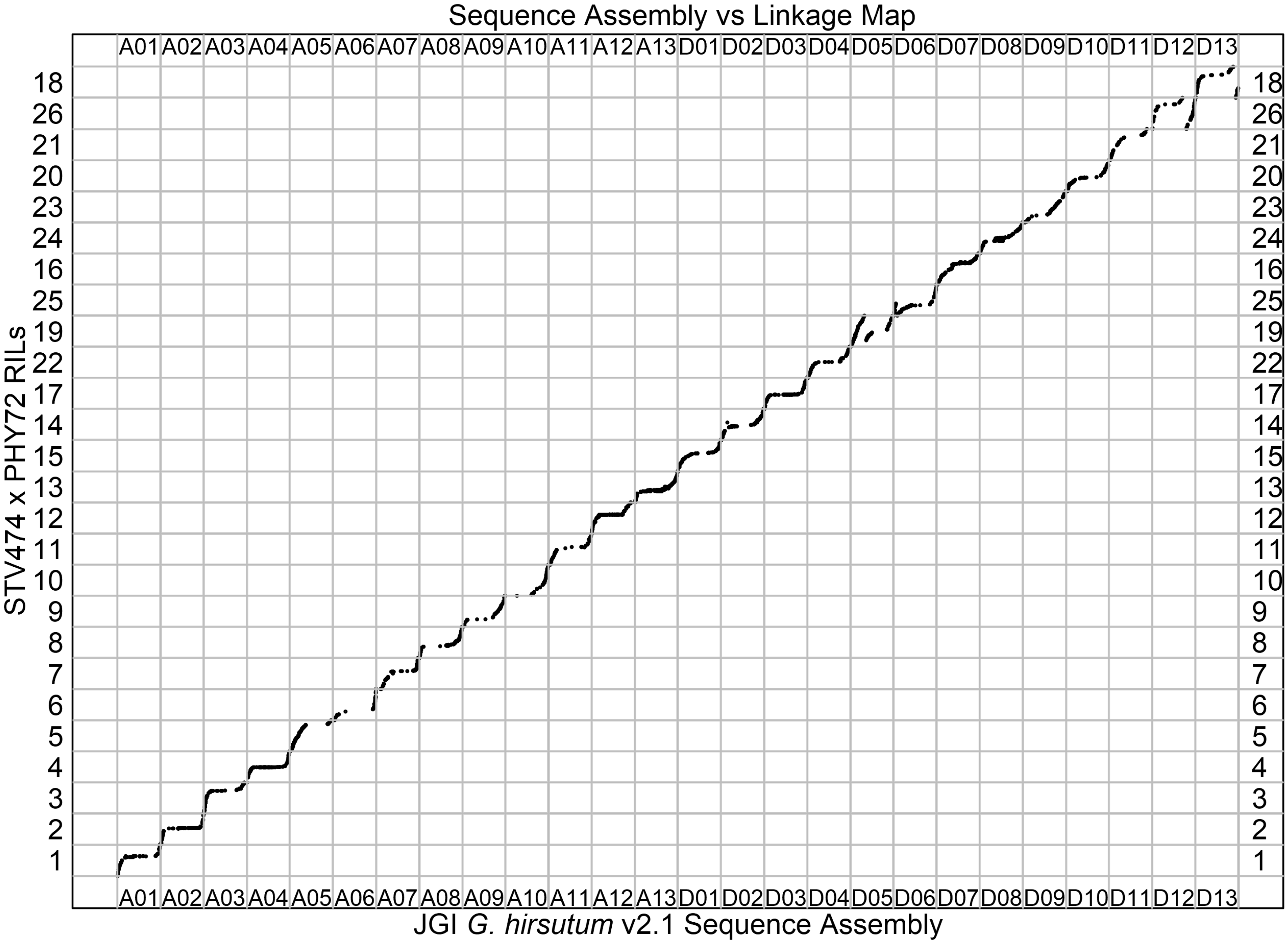

Figure 2.4 (continued)

I.

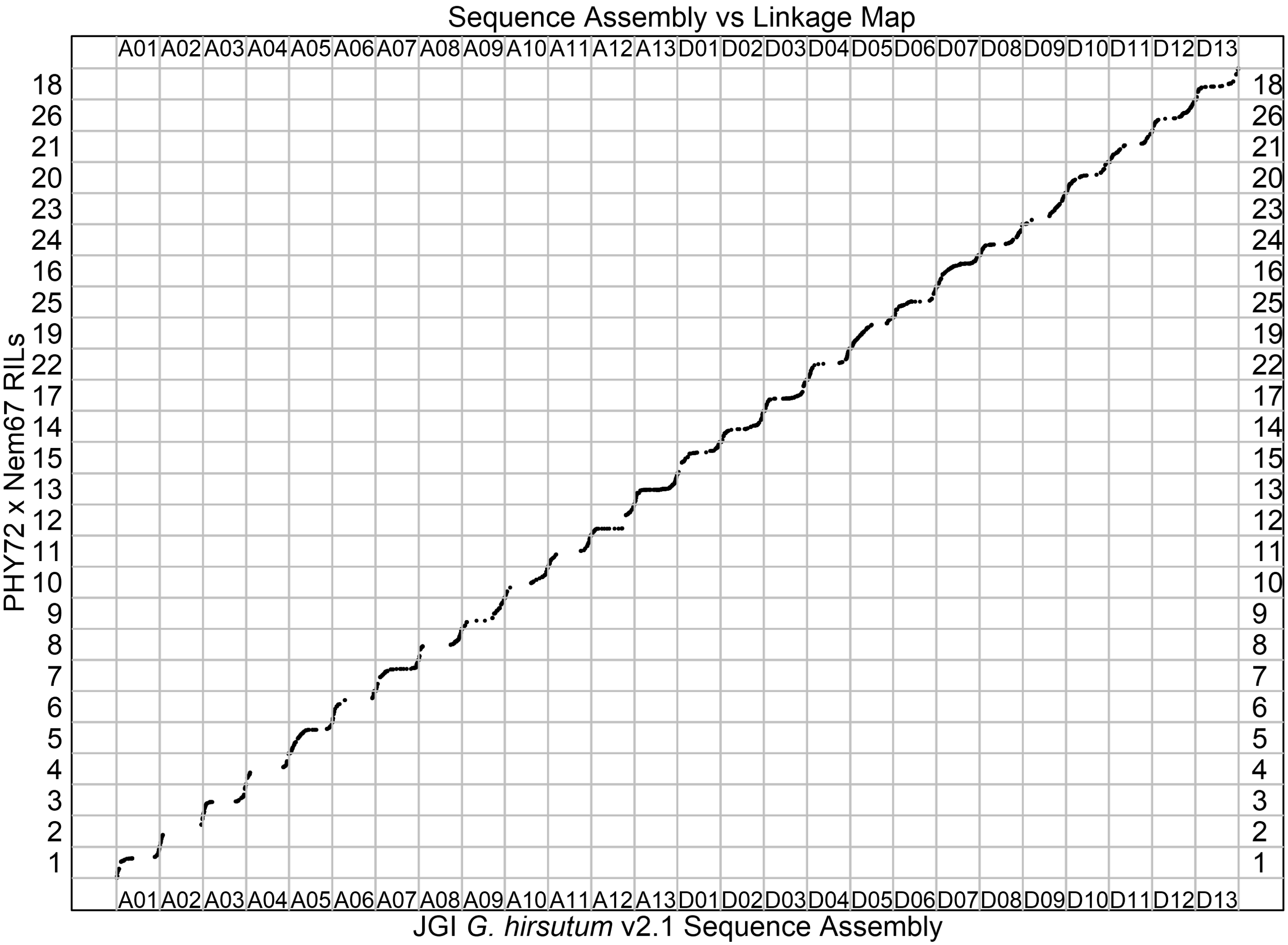

Figure 2.4 (continued

Haplotype Structure of Improved and Wild Cotton Accessions

A total of 19,856 and 22,381 SNP markers aligned to the JGI *G. hirsutum* v2.1 sequence assembly were used to delimit haplotype blocks within the 257 improved and 71 wild cotton accessions, respectively. Using PLINK's CI the haplotype block demarcation approach, 2,064 and 2,503 genome-wide haplotype blocks were found within the improved and wild cotton panels, respectively (**Figure 2.5**). The average haplotype block length in the wild cotton panel was approximately 64% the size of the average haplotype block length in the improved cotton panel and the distribution in block sizes in both cotton panels ranged up to 75,378 kb (**Table 2.2, Table 2.3,** & **Figure 2.6**). Similarly, the improved cotton panel was found to contain a similarly sized haplotype block of 72,350.1 kb in length. In both the two germplasm pools, the largest haplotype block was in chromosome A08.

The median number of haplotypes per haplotype block within both cotton panels were estimated by HaploView to be 3.0 while the average number of haplotypes per haplotype block were estimated to be 3.3 and 4.0 in the improved and wild cotton panels, respectively (**Table 2.2**). There is a non-linear, positive correlation between the number of haplotypes per haplotype block length with a greater linear correlation in the improved cotton panel compared to the wild cotton panel (**Figure 2.7**). The wild cotton panel contains a significantly greater number of haplotypes per haplotype block (**Figure 2.8B**) as well as significantly shorter haplotype blocks (**Figure 2.8A**).

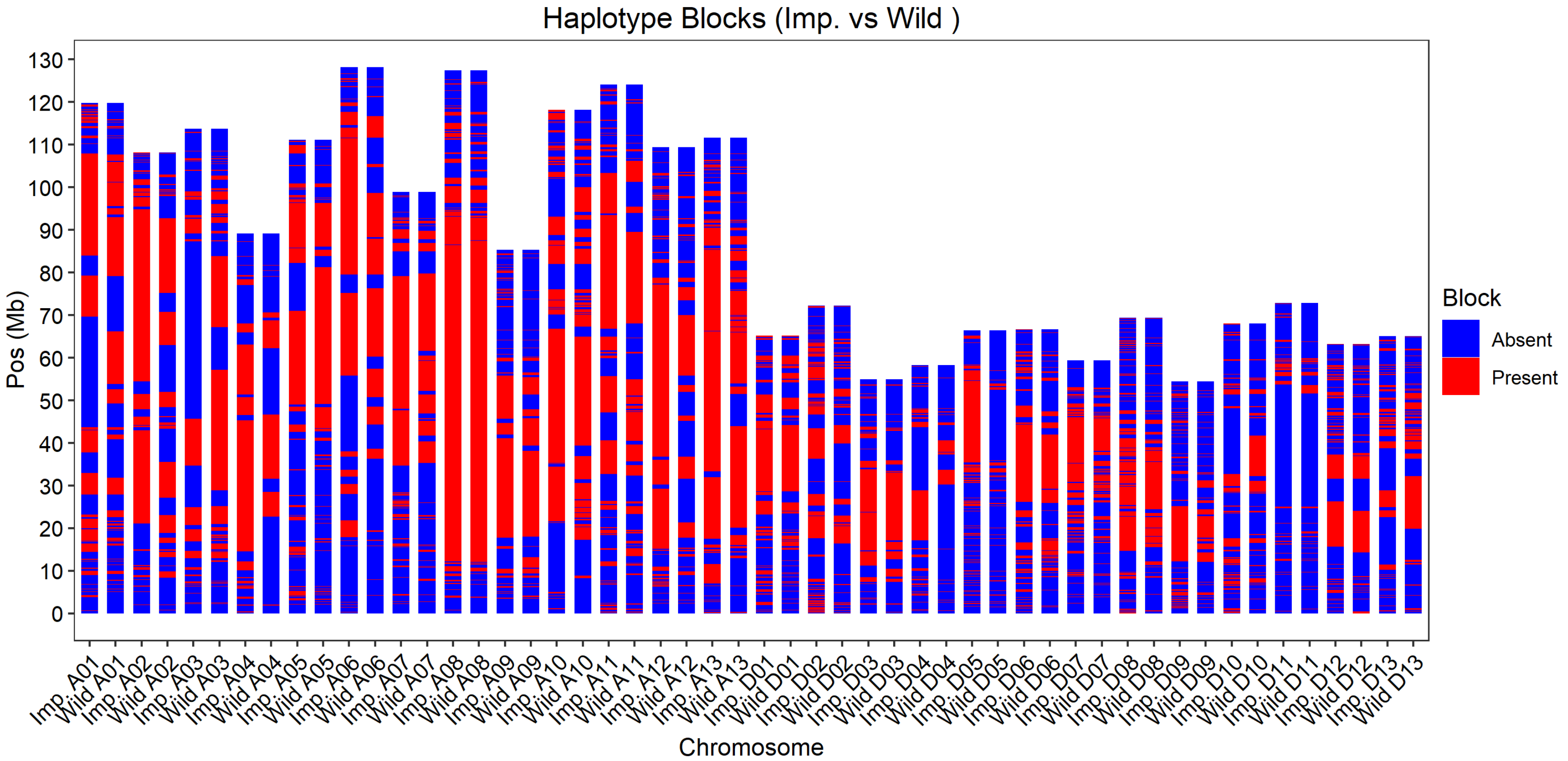

Figure 2.5 The haplotype structure of the improved and wild cotton panels. Regions within haplotype blocks are shown in red while regions outside of haplotype blocks are shown in blue. Chromosomes (A1-A12 and D1-D12) are represented by the x-axis while position (Mb) is represented by the y-axis. Improved and wild chromosomes are shown in pairs.

Table 2.2 Summary statistics on the haplotype structure of the improved and wild cotton panels.

	Statistic	Improved	Wild
Haplotype Blocks	Mean	600.41 kb	386.25 kb
	Min	0.002 kb	0.002 kb
	Max	72,350.1 kb	75,378 kb
	Median	29.42 kb	12.25 kb
	Block Count	2,064	2,503
	Block Coverage	1,239.25 Mb	966.81 Mb
Number of Haplotypes per Block	Mean	3.3	4
	Min	2	2
	Max	22	25
	Median	3	3
	Haplotype Count	6,923	10,050

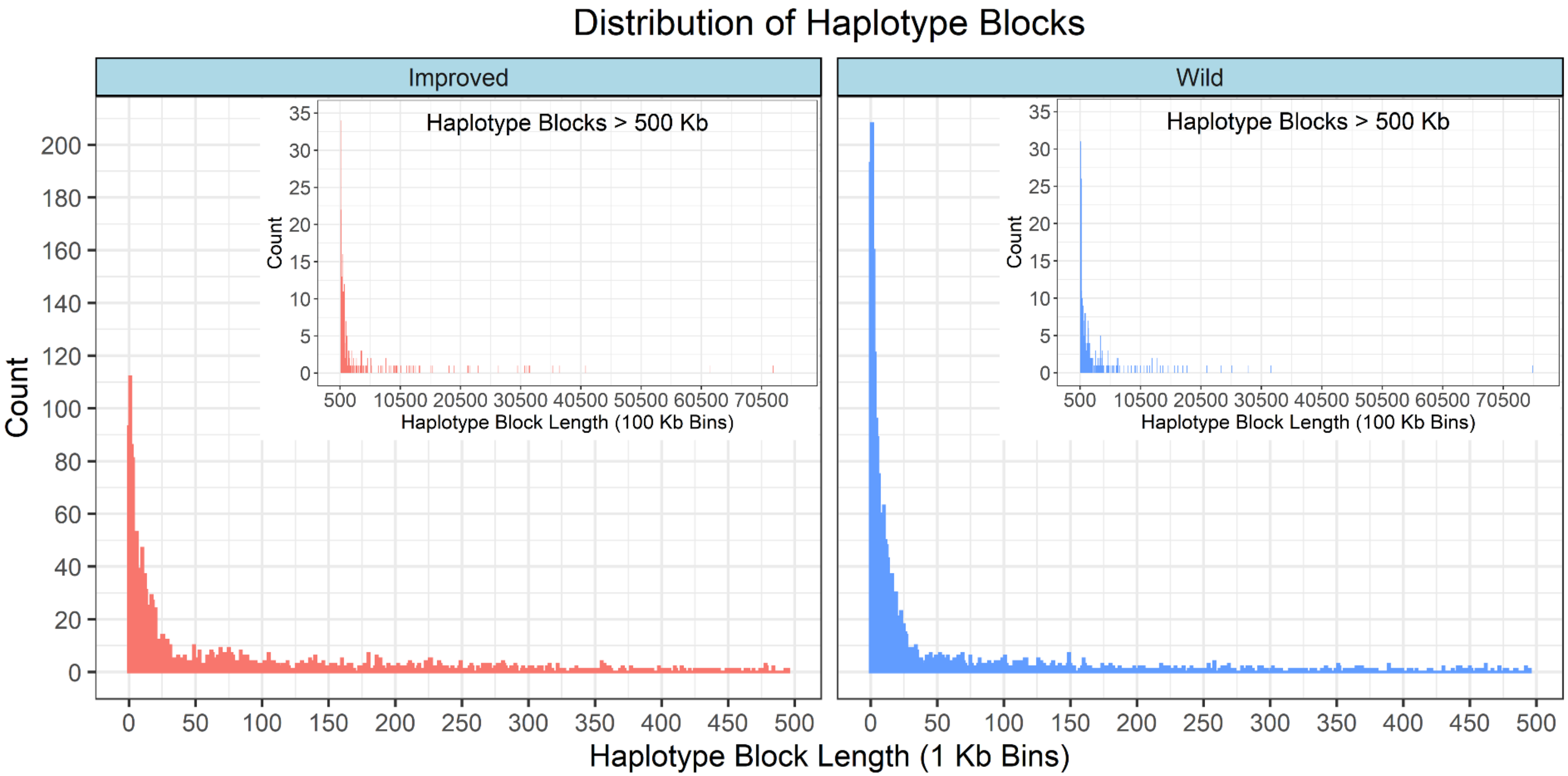

Figure 2.6 The distribution of haplotype block length in improved and wild cotton. The improved cotton panel (red) is shown on the left while the wild cotton panel (blue) is shown on the right. The length of haplotype length is represented by the x-axis while the number of haplotypes of a given length is represented by the y-axis. Upper-right distribution represent blocks greater than 500 kilobases.

Table. 2.3 The percentile distribution of haplotype block length within the improved and wild cotton panels.

Percentile	Improved (Kb)	Wild (Kb)
5%	0.56225	0.2851
10%	1.5226	0.6904
15%	2.7068	1.3376
20%	4.2008	2.0244
25%	6.441	3.025
30%	9.4713	4.075
35%	11.9463	5.724
40%	15.966	7.466
45%	20.182	9.933
50%	29.416	12.253
55%	52.8815	15.6281
60%	74.4892	19.6458
65%	100.24295	27.124
70%	140.9213	49.4876
75%	196.46075	81.9725
80%	266.0568	131.604
85%	361.0734	233.5109
90%	598.3836	458.9022
95%	1,259.8865	1,310.284

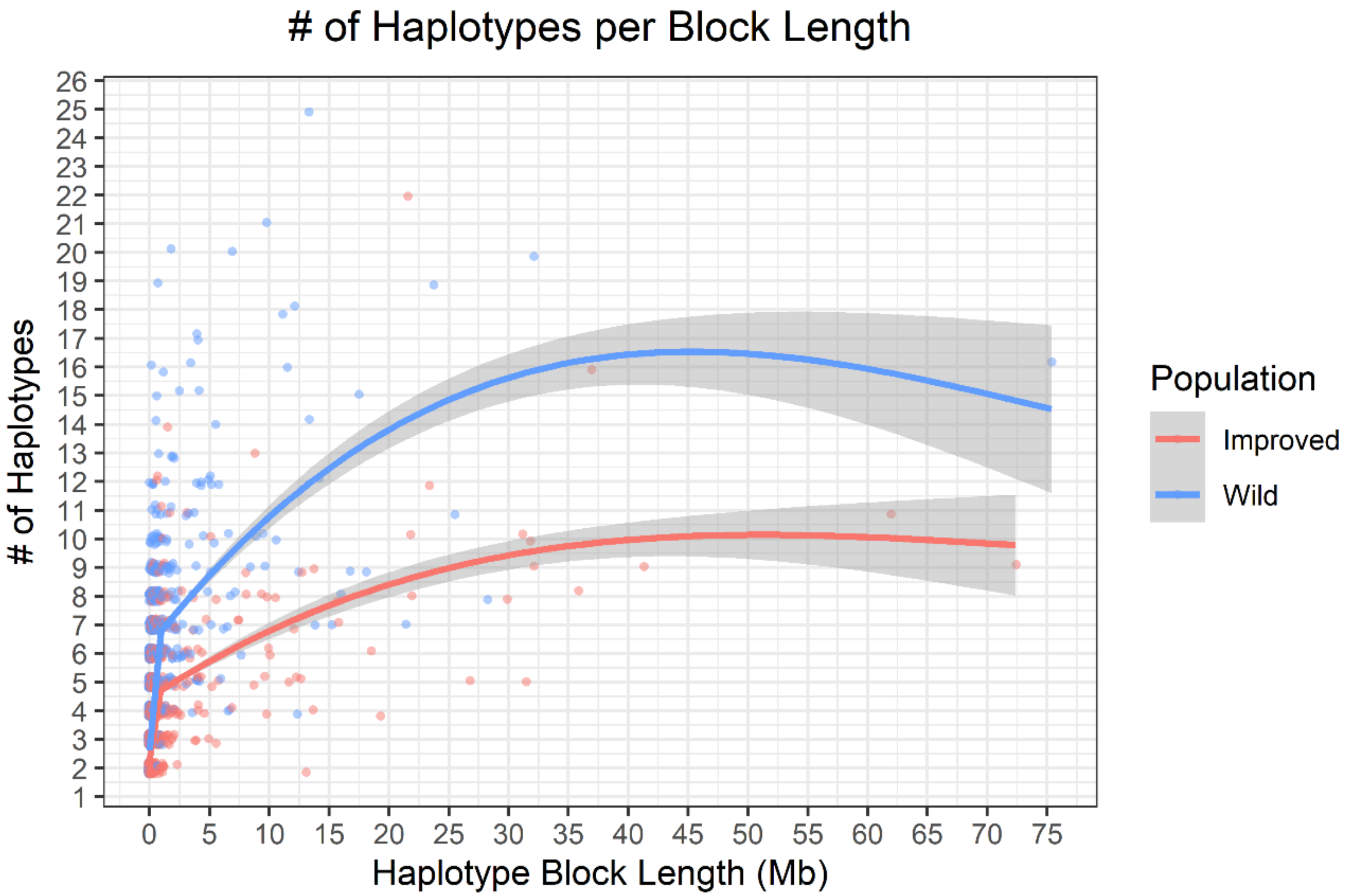

Figure 2.7 The number of haplotypes per haplotype block in panels of improved (n=257) and wild cotton (n=71). Circles in red represent improved cotton haplotype counts while circles in blue represent wild cotton haplotype counts. The haplotype length is represented by the x-axis while the number of haplotypes per haplotype block is represented by the y-axis. Non-linear regressions shown were conducted using a general additive model (GAM).

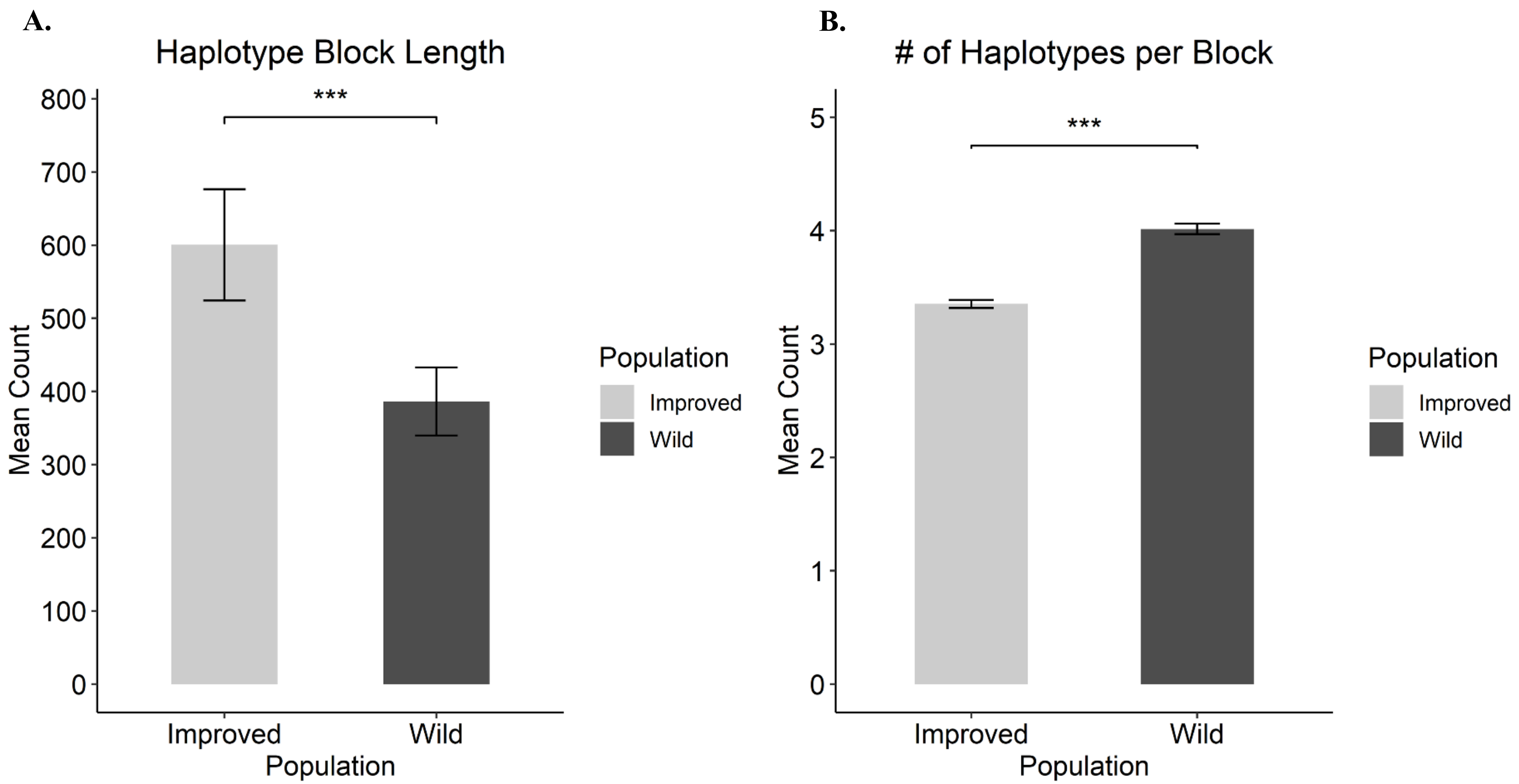

Figure 2.8 A) The average haplotype block length and the B) average haplotype count per haplotype block between the improved and wild cotton panels. Statistical test was conducted using a Wilcoxon rank-sum/Mann-Whitney unpaired test (p-value < 0.001).

Recombination Rate Estimations and the Haplotype Structure of Chromosome A08

Three regression models, 1 non-linear and 2 linear, were used to estimate the genome-wide recombination rates for all 9 mapping populations as a function of genetic distance (cM) over physical distance (Mb) (**Figure 2.9**). With the exception of linkage groups 18, 19, 25, and 26 of the STV474 x PHY72 RIL population, recombination rates within all individual linkage groups within each population were estimated (**Figure 2.10**). Both non-linear regression models resulted in negative interpolations that were manually set to values of zero. The cubic spline with generalized cross validation (CS-CV) resulted in the widest distribution of recombination rates with values greater than 90 cM/Mb occurring. The LOESS and sliding window resulted in narrower distributions with similar regression values. The LOESS regression was ultimately used for comparing the relationship between recombination rates and the haplotype structure of chromosome A08 due to its interpolation across genomic areas containing low SNP densities allowing for uninterrupted chromosome-wide regression estimation.

Chromosome A08 was partitioned into 69 haplotype blocks ranging from 0.145 kb to 72,350.1 kb. The range in the number of haplotypes per haplotype block was 2 to 9 with the greatest number of haplotypes found in the ~ 72 Mb haplotype block. A total of 1,799 SNPs were used to generate the haplotype structure of chromosome A08 of which 1,070 SNPs were used to generate the ~72 Mb haplotype block.

The recombination rates estimated by the LOESS regression of chromosome A08 (**Figure 2.11A**) were superimposed to the haplotype structure of chromosome A08 of the improved cotton panel (**Figure 2.11B**). The haplotype structure of chromosome A08 for the improved cotton panel were generated using CIs and visualized using HaploView. Haplotype blocks were visually demarcated by black triangles. Linkage disequilibrium (*D'*) was measured as a value ranging from 0 (blue) to 1 (red) where a value of '0' indicates 0% LD between SNPs while a value of '1' indicates 100% LD between SNPs with no recombination.

Validity of the ~ 72 Mb haplotype block in chromosome A08 was confirmed by sequencing a small panel of improved cotton accessions using a target-capture approach through paired-end sequencing. A total of 22 individuals and 9,875 SNPs were retained following quality control. A total of 766 SNPs were found on chromosome A08 and were used to delimit 21 haplotype blocks (**Figure 2.12**). The largest haplotype block was 80,816.9 kb in size contained 539 SNPs. This block was homologous in location to the large ~72 Mb block found in the improved cotton panel.

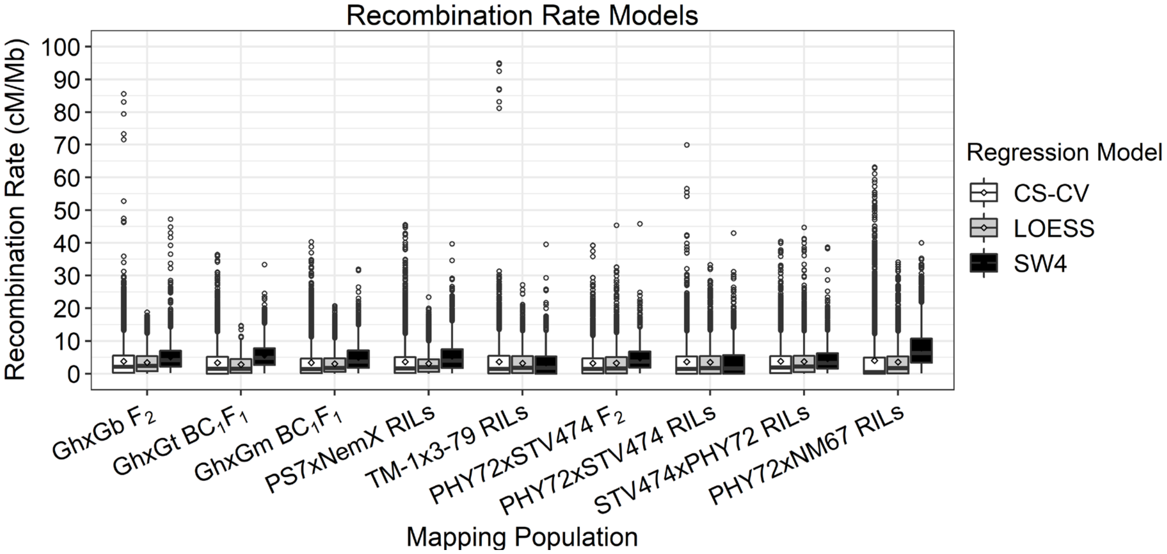

Figure 2.9 Distribution of recombination rate values for three recombination models. Recombination rates for a cubic spline with cross-validation (CS-CV), a locally estimated scatterplot smoothing (LOESS), and a non-overlapping 1 Mb sliding window (SW4) regression are shown. All 9 mapping populations are shown on the x-axis and recombination rate (cM/Mb) is represented on the y-axis.

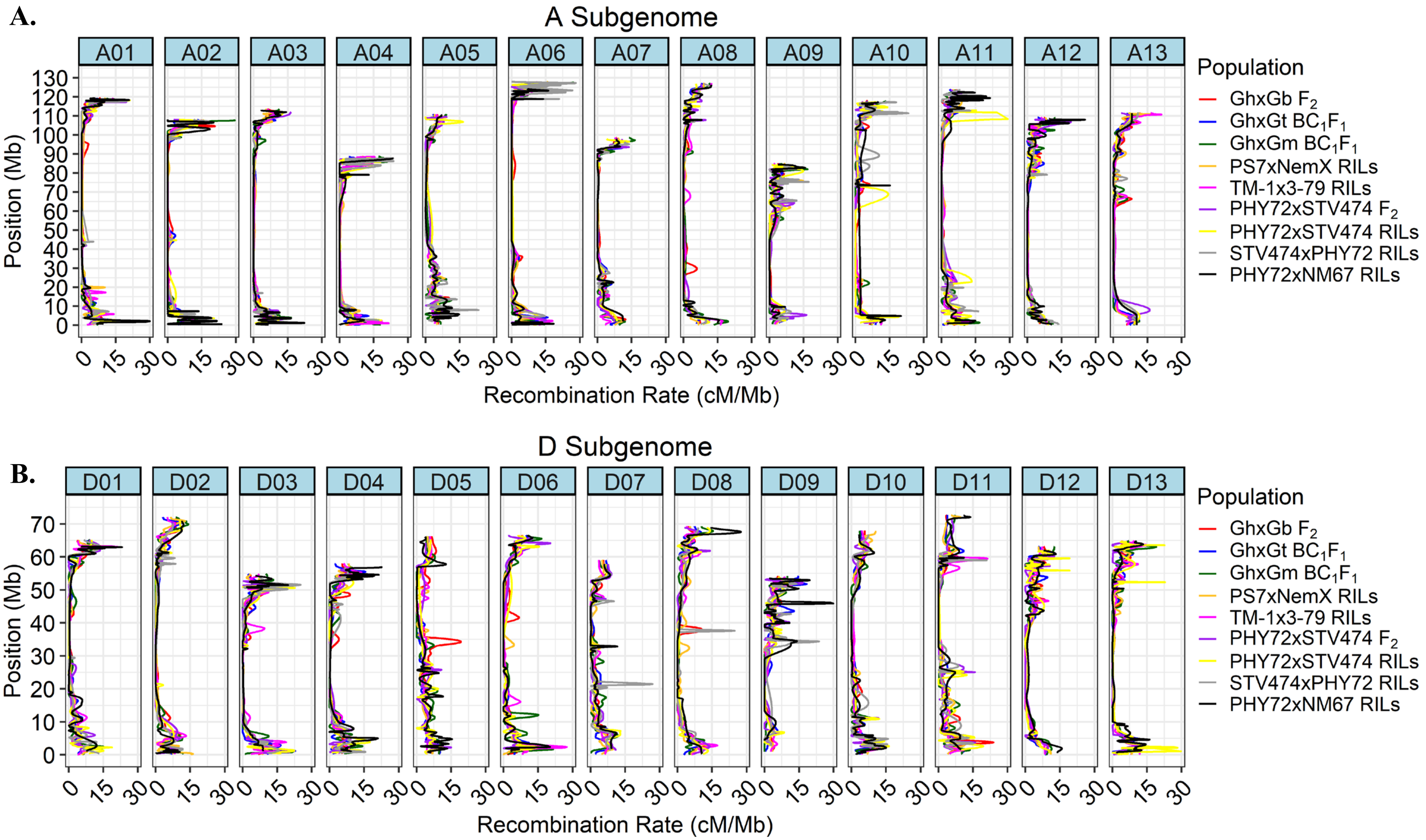

Figure 2.10 The recombination rate maps for nine mapping populations. The A) A subgenome is represented on the top panel while the B) D subgenome is represented on the bottom panel.

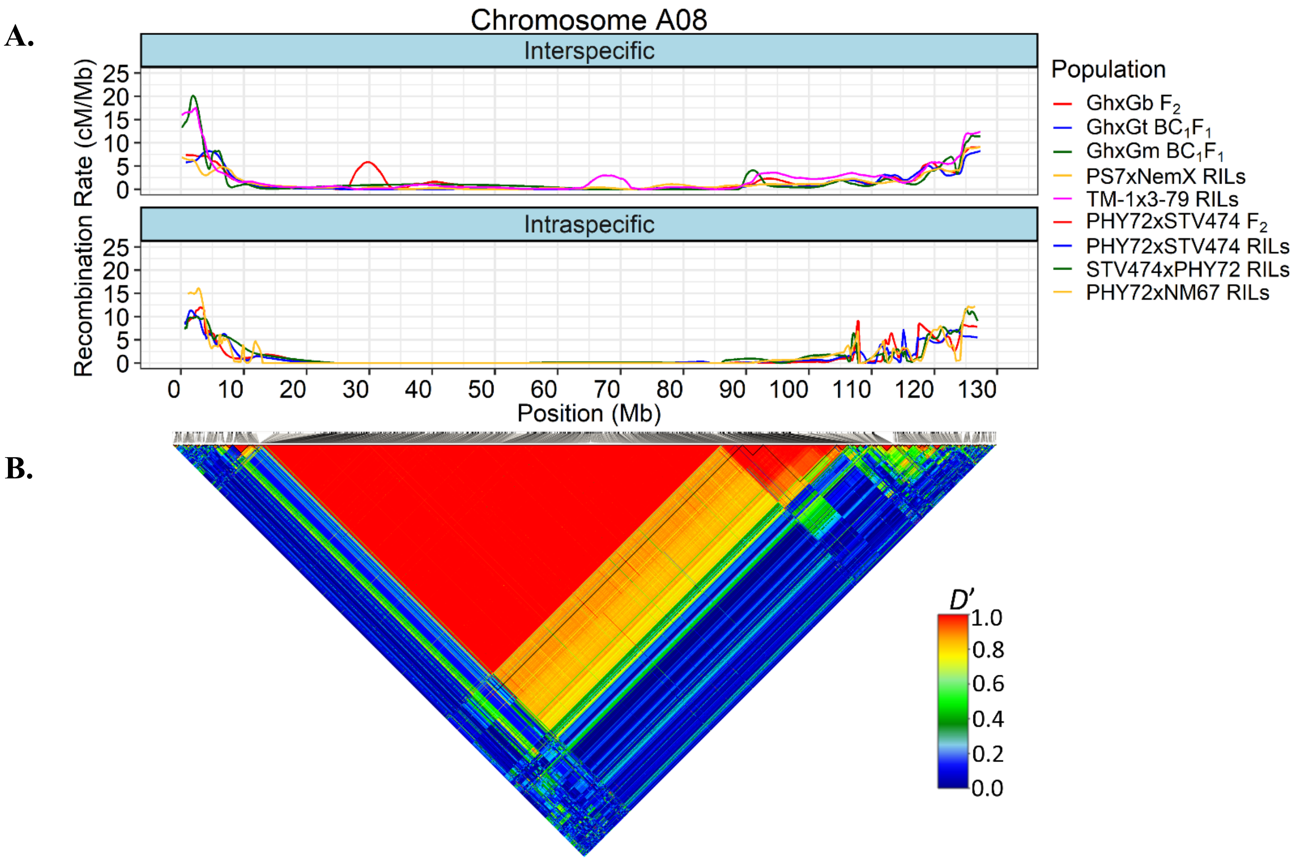

Figure 2.11 Chromosome-wide recombination map of nine mapping populations and haplotype structure of the improved cotton panel for chromosome A08. A) The five interspecific population are shown on the top panel while four intraspecific populations are shown on the bottom panel. Marker position (Mb) is displayed on the x-axis and the recombination rate (cM/Mb) as a function of genetic distance over physical distance is displayed on the y-axis for the upper panel. **B)** Heatmap consists of equidistant tiles that indicate linkage disequilibrium as determined by a normalized coefficient of linkage disequilibrium (D') between pairs of markers. Black triangles within the heatmap demarcate haplotype boundaries Markers corresponding to SNP positions above the heatmap are congruent to the x-axis of the recombination map.

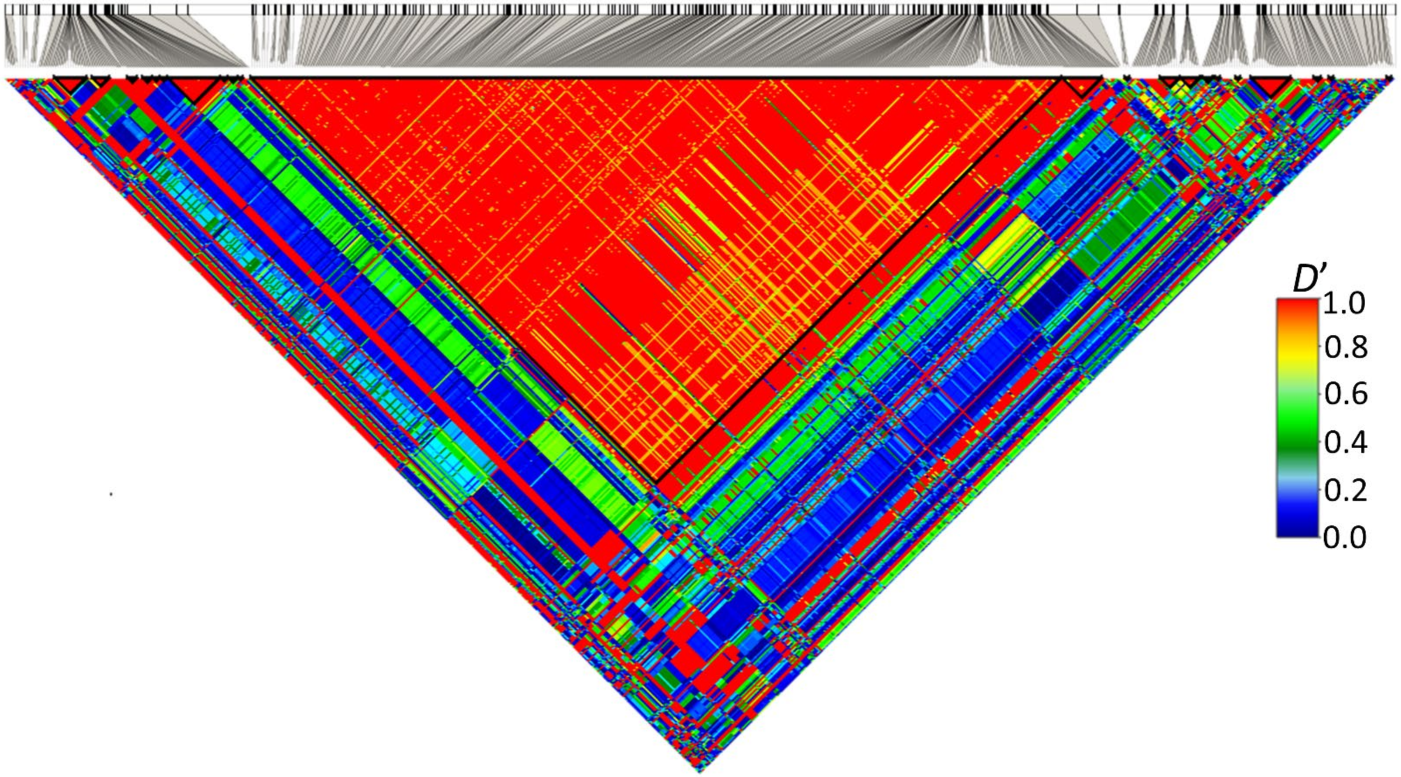

Figure 2.12 The haplotype structure of chromosome A08 in 22 improved *G. hirsutum* accessions via paired end sequencing. Heatmap consists of equidistant tiles that indicate linkage disequilibrium as determined by a normalized coefficient of linkage disequilibrium (D') between pairs of markers. A total of 539 SNPs were used to generate the haplotype structure. Haplotype blocks are delimited by black triangles within the heatmap.

Selective Sweeps and Haplotype Block Expansion

Unprocessed genotype data of the improved and wild cotton panels were pooled together and filtered for quality to generate a meta-population. A total of 20,814 SNPs with an average density of 9.12 SNPs/Mb were found between the improved and wild panels and were used to detect selective sweeps between both populations (**Figure 2.13**). The number of haplotype clusters, denoted as *K*, within the metapopulation was estimated to be 17 due to this *K* value resulting in the lowest cross-validation error when performing population clustering **(Figure 2.14)**. A *K* value of 17 was then used to determine haplotype clustering when scanning for selective sweeps using hapFLK. A total of 1,373 SNPs were found to be statistically significant (p-value < 0.05) prior to FDR correction and 105 SNPs were found to be statistically significant (q-value < 0.05) (**Figure 2.15A**) following for FDR correction (**Figure 2.15B**). The 17 haplotype cluster frequencies for chromosome 25 (D08) estimated by hapFLK were (**Figure 2.16B-C**) compared to the selective sweeps between the improved and wild cotton panels on chromosome 25 (**Figure 2.16A**) as well as to their haplotype structures as estimated by CI (**Figure 2.17A-C**).

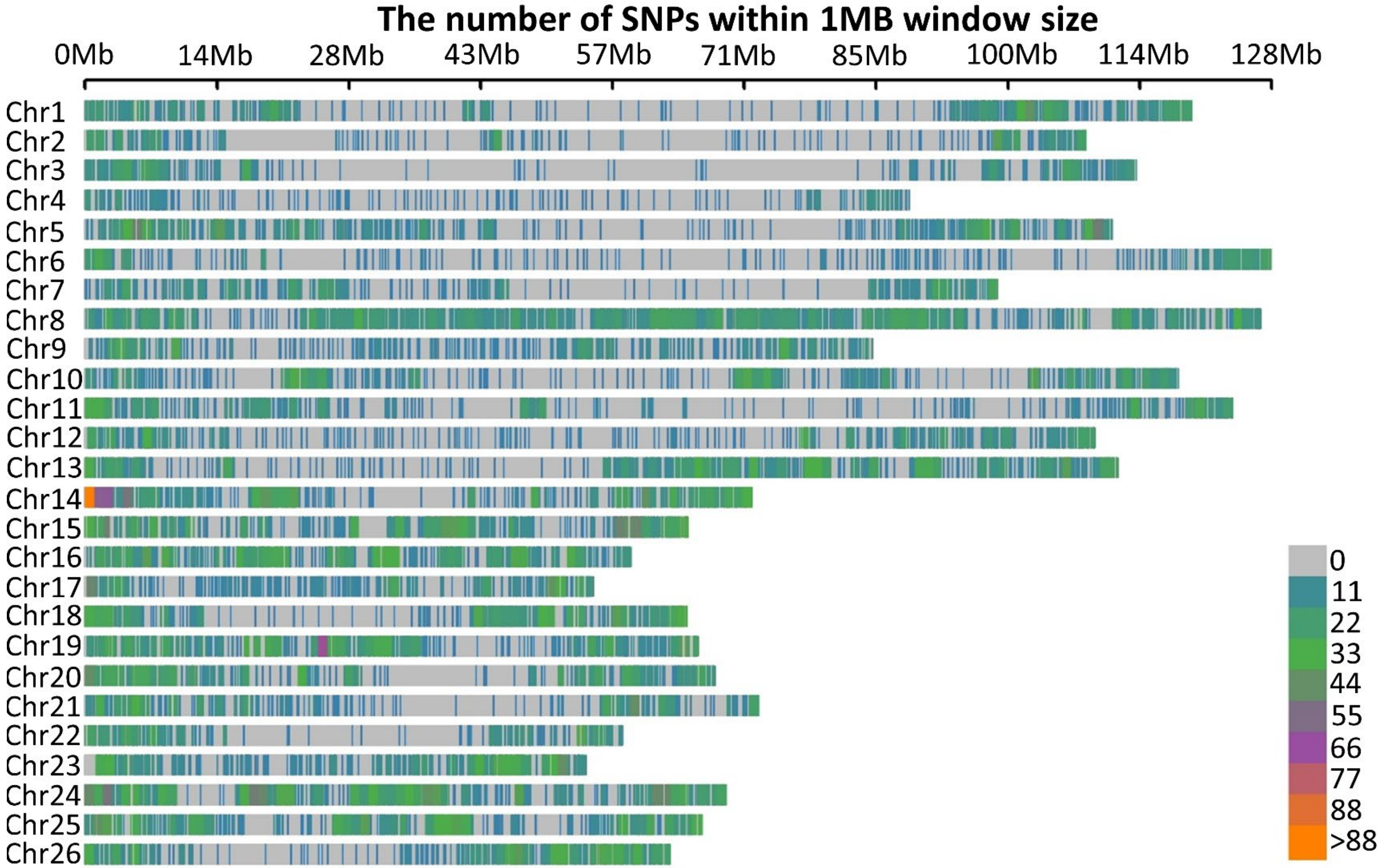

Figure 2.13 SNP density within 1 Mb windows of a metapopulation consisting of 328 improved and wild cotton individuals. Position of SNPs (Mb) is shown on the x-axis while the chromosomes identities (1-26) are shown on the y-axis.

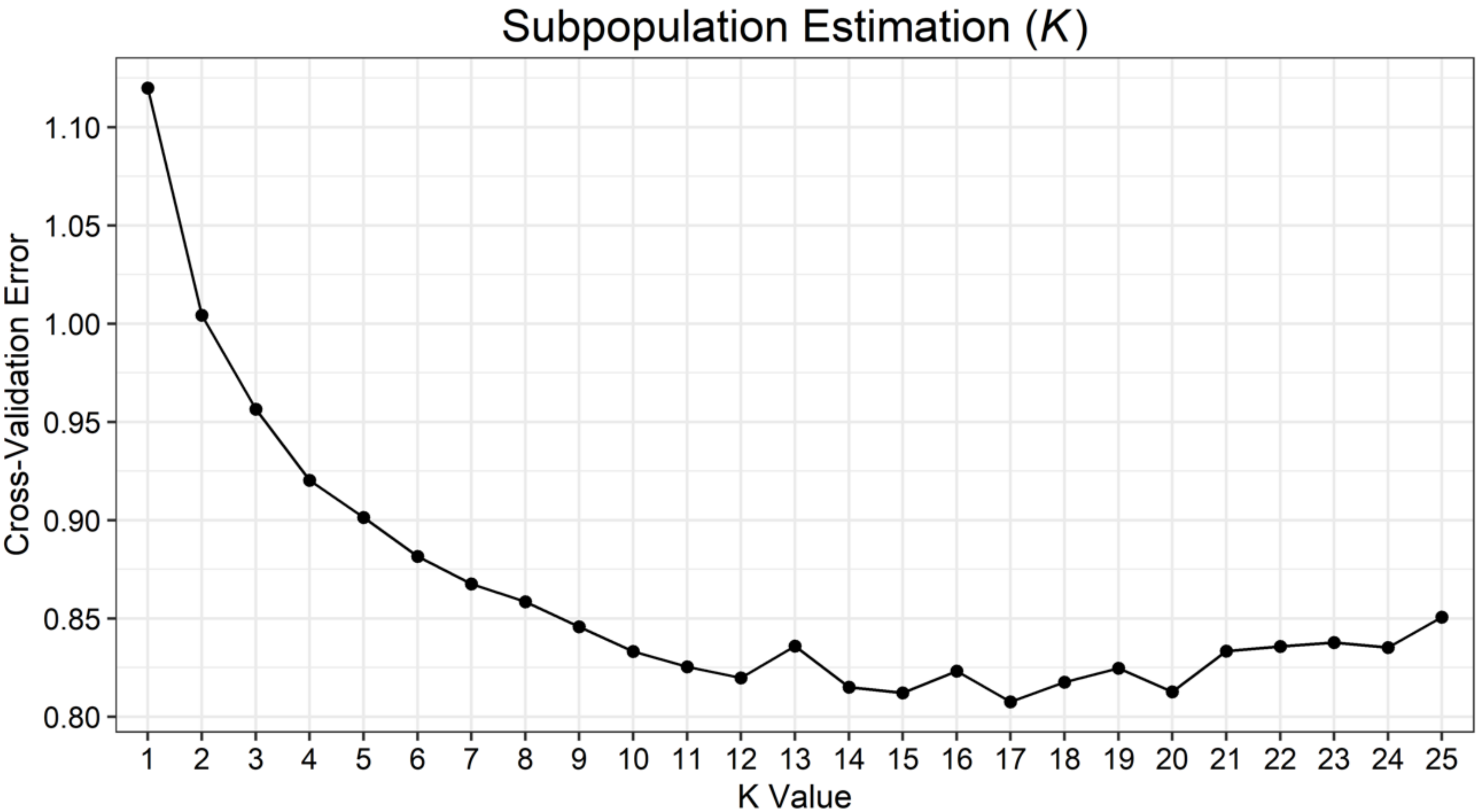

Figure 2.14 Cross validation analysis for each number of subpopulations modeled. Lower K value indicates better modeling. Lowest K indicates the best estimate for the number of subpopulations.

A.

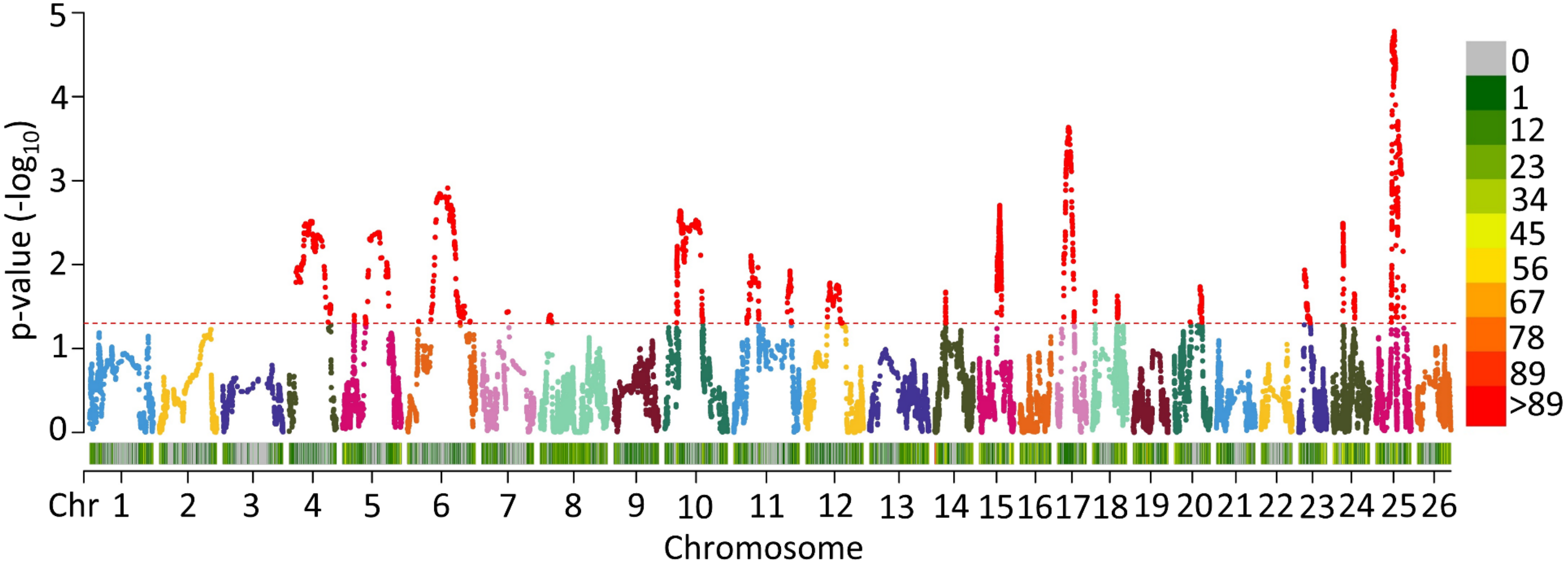

Figure 2.15 Manhattan plot of the number of statistically significant peaks. A) Uncorrected values (p-value < 0.05) are illustrated on the top while **B)** values following multiple test correction (FDR; q-value < 0.05) are illustrated on the bottom.

B.

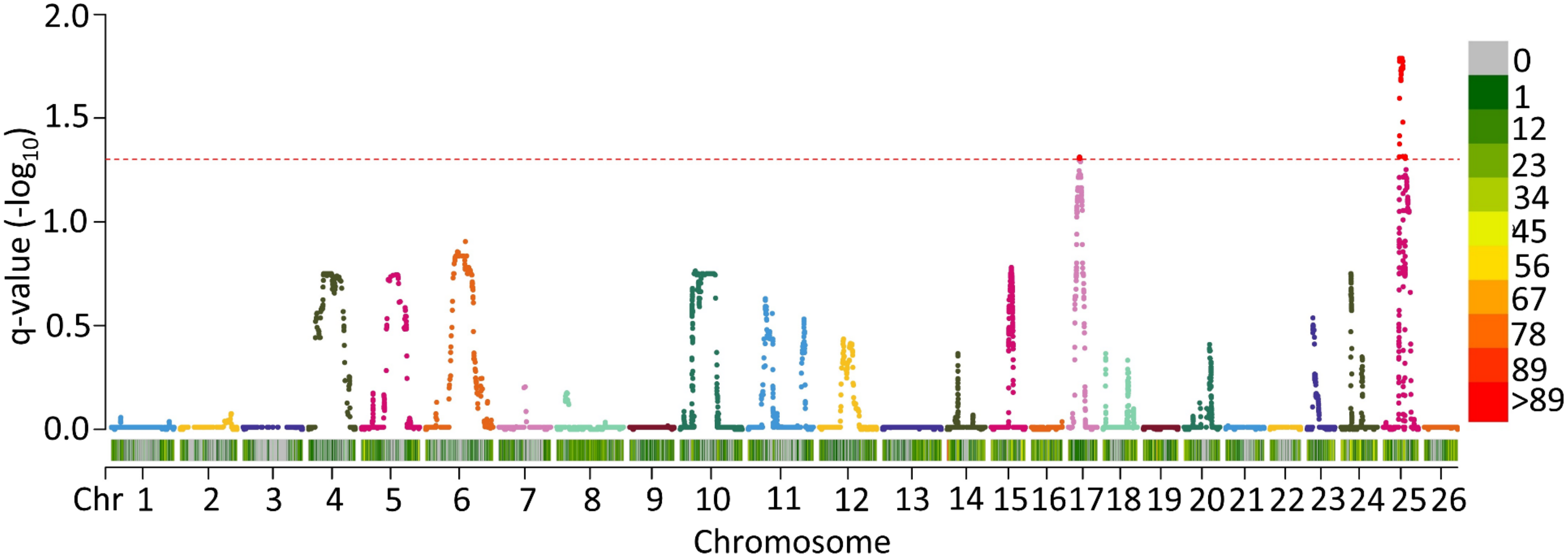

Figure 2.15 (continued)

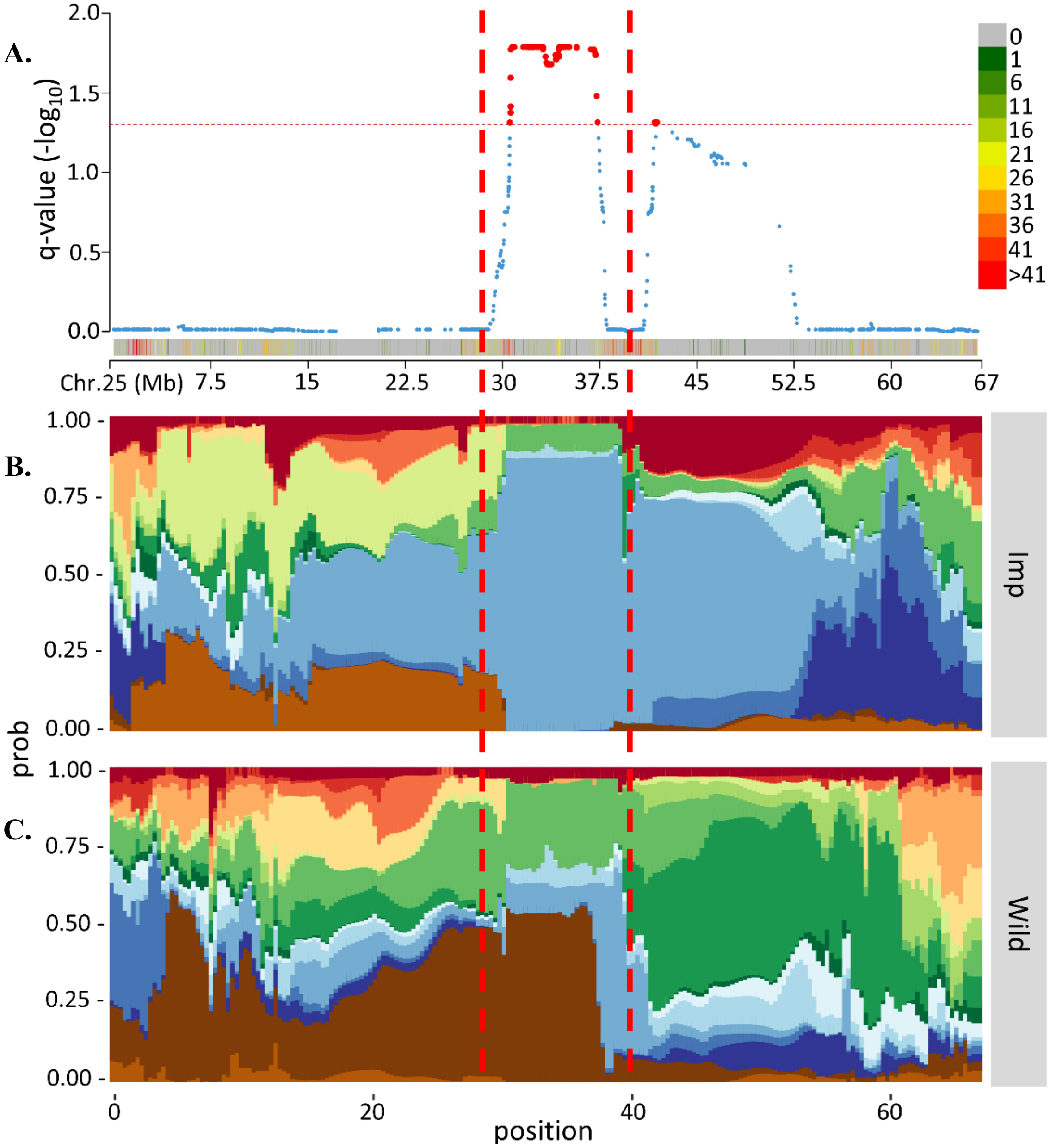

Figure 2.16 Cluster plot representing the haplotypes in Chromosome 25. A selection sweep of a certain haplotype (regions between dashed red lines) is seen between 30 Mb and 44 Mb. **A)** Selective sweep and the corresponding haplotype cluster frequencies for the **B)** improved and **C)** wild cotton panels is shown. Each color represents a haplotype grouped by K values.

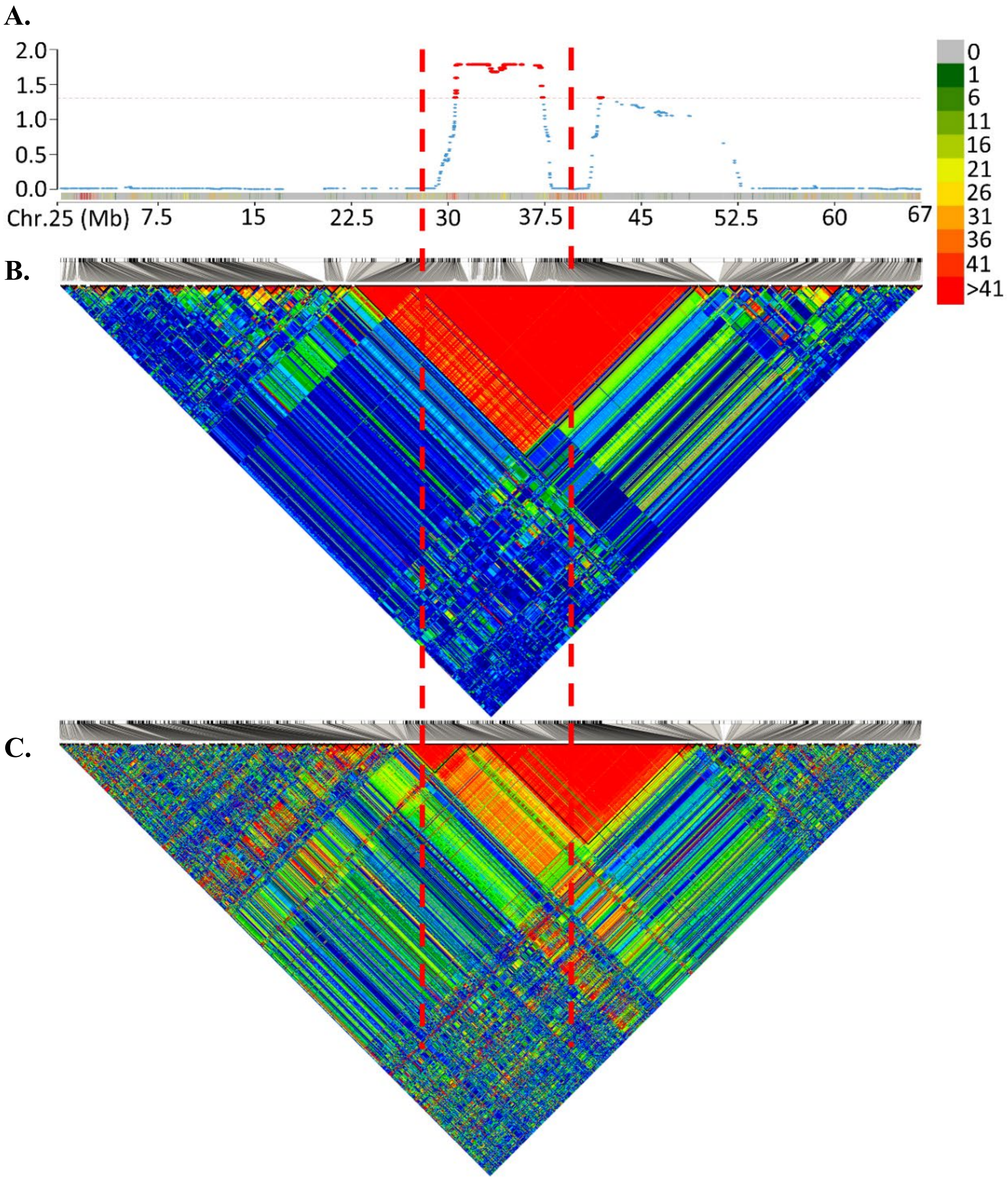

Figure 2.17 The haplotype structure of chromosome 25 and its selective sweep. A) The selective sweep (q-value <0.05) indicated between red dashes and its corresponding haplotype structure for **B)** the improved and **C)** wild cotton panels.

Discussion

A major challenge to variant calling in *G. hirsutum* arises from the inability to distinguish homeologous SNPs from homologous SNPs that arise due to sequence similarity between the two subgenomes. A second challenge is the ambiguity of mapping oligos containing known SNPs to the correct regions within a genome. This is due to the highly repetitive nature of DNA sequences between homeologs and within homologs, including remnants of infectious elements and vestiges of repeated ancient polyploidization events before and after formation of the basic *n*=13 *Gossypium* genome [6]. In this study, homeo-SNPs were addressed using the methodology described in Hulse-Kemp et al. (2015) and ambiguous alignment of putative homeologous oligos were resolved using mapping populations. The latter was accomplished by taking advantage of the nature of recombination fractions between markers in which markers that are closer together will share similar recombination fractions compared to markers that are farther apart on a chromosome. Thus, the recombination fractions of markers allow for their grouping and linear arrangement. This is the basis of linkage mapping which was first hypothesized in 1911 [85] and experimentally demonstrated in 1913 [86]. In this study, a total of 19,856 SNPs were mapped in the improved cotton panel and 22,381 SNPs were mapped in the wild cotton panel (**Table 2.1**). Out of the total number of mapped SNPs, 19,829 and 22,350 SNPs within the improved and wild cotton panels, respectively, were also mapped in the 5 interspecific and 4 intraspecific mapping populations (**Figure 2.3**). A total of 2,695 SNPs (13.6%) and 2,820 (12.6%) SNPs in the improved and wild panels originally aligned to either a homeolog or non-homologous chromosome and were corrected for using the mapping locations provided by the linkage mapping populations. The correction of ambiguously mapped SNPs was essential to accuracy of haplotype structure estimation for *G. hirsutum*, especially for the larger haplotype blocks because they tend to contain greater numbers of SNPs. Thus, they suffer increased probabilities of SNPs from other areas being erroneously mapped to them and/or some of their SNPs being incorrectly genotyped. Differential sensitivity of larger haplotype blocks to ambiguously assigned SNPs reflects their propensity to be incorrectly partitioned, because they are more likely to suffer from localized erroneous SNP genotyping that exceeds CI threshold bounds for block partitioning.

The introgression of beneficial interspecific germplasm into cultivated cotton can be constrained by a lack of meiotic recombination between homologous chromosomes, and/or the absence of transmission of recombinants. This can prevent or slow the separation of advantageous and deleterious alleles of neighboring donor loci, preclude or impede the creation of segregates with maximum numbers of levels of beneficial traits and no critically deleterious ones, which in turn hinders breeding efforts.

The lack of meiotic recombination can be measured as a decrease in the non-random association of genetic markers along a chromosome, which is also known as linkage disequilibrium. Linkage between adjacent markers generally occurs due to physical proximity, and continuous linkage of markers that are consistently co-inherited as a unit form linkage blocks, or haplotype blocks. Haplotype blocks thus are a basis for linkage drag and prevent or slow the generation of novel allelic combinations. The characterization of the haplotype structure of improved and wild cotton representative of the current genetic diversity of *G. hirsutum* provides a snapshot of genomic areas prone to linkage drag or under heavy selection (**Figure 2.5**).

The characterization of the haplotype structure of improved cotton using the Cotton63KSNP Array revealed that of 5% haplotype blocks are equal to or less than 0.0 56225 Kb, 50% of are equal to or less than 29.416 Kb, and 75% are equal to or less than 196.46075 Kb (**Table 2.3**). The total size of the genome encompassed within haplotype blocks was estimated to be 1,239.25 Mb out of a total genome size of 2,305.24 Mb (53.76 %) within the improved cotton panel and 966.81 Mb (41.94%) within the wild cotton panel. The size of the haplotype blocks within the wild cotton population are relatively half of those of the improved cotton population (**Figure2.8A**). This is likely due to the genetic bottlenecks created by the commercialization of cotton following post-Columbian cultivation in which increased selection for fiber quality led to large homogenization of cultivated cotton. The average number of haplotypes per haplotype block in improved cotton is also significantly less than those in wild cotton (**Figure2.8B**). Although this is not inconsistent with a lack of genetic diversity, it is surprising, perhaps even alarming, given the four-fold difference in population sizes used in this study, where the improved cotton population comprised 271 individuals while the wild cotton population comprised 71 individuals, yet the average number of haplotypes was significantly lower in the larger (improved) population. Regardless of the comparison, the average number of haplotypes per block in improved cotton is 3.3, a value that upon evaluating is extraordinarily low. Collectively encompassed by these statistics are approximately 24,270 (32.2 %) of the recognized 75,376 genes of the *G.*

hirsutum genome assembly [44]. Most importantly, this analysis provides critical detail and insight into the distribution and severity of genetic uniformity and limitations in cultivated cotton, related non-domesticated resources. Both seem to beg for diversification, e.g., through increased recombination and interspecific introgression.

The largest haplotype block in both improved and wild populations were found on chromosome A08. This block is approximately 72.4 Mb in improved cotton and 75.4 Mb in wild cotton and its presence in both populations suggests it is pre-Columbian in origin (**Table 2.1**). There are 1,070 SNP markers distributed across the haplotype block in improved cotton and 1,101 SNP markers in wild cotton, which implies that the large haplotype block is not due to lack of sufficient SNP density. The recombination landscapes of the 5 interspecific and 4 intraspecific mapping populations also exhibit recombination suppression within the region that is homologous to the large haplotype block in improved and wild cotton with two notable exceptions (**Figure 2.11**). The *G. hirsutum* x *G. barbadense* F_2 and *G. hirsutum* x *G. barbadense* RIL both reveal recombination events at approximately 30 Mb and 68 Mb, respectively. These results imply that it is theoretically feasible to introgress interspecific germplasm from *G. barbadense* into the large haplotype block of *G. hirsutum* at those specific regions and provides evidence for the disruption of haplotype blocks through wild germplasm introgression. The generation of multiple interspecific and intraspecific recombination maps (**Figure 2.10**) in this study also provides breeders with the ability to utilize marker-assisted selection to leverage the differential recombination landscapes between intraspecific and interspecific populations to separate traits under linkage drag (**Figure 2.1** & **Figure2.10**).

In this study, putative selective sweeps between improved and wild *G. hirsutum* were used as a proxy for domestication. These putative results were compared to the haplotype structure of a region under significant selection to characterize the effects of domestication on the frequency of haplotypes. A strong selection sweep on chromosome 25 (D06) was detected between 30 Mb and 37 Mb and between 41 Mb to 42 Mb (q-value < 0.05; Figure 2.14). Previous genome wide association studies (GWAS) and quantitative trait loci (QTL) analysis reveal fiber-related traits associated to adjacent selective sweeps on chromosome 25 [87, 88]. The haplotype frequencies within the significant selective sweeps also display an increase homogenization implying a decrease in haplotype diversity at those regions (**Figure 2.16B-C**). When compared to the haplotype structure of chromosome 25 of improved and wild cotton, an increase in the haplotype block length is seen in improved cotton relative to wild cotton at the putative region under the selective sweep (**Figure 2.17B-C**). This provides evidence that the expansion of haplotype blocks can occur due to genetic bottlenecks which results in the decrease of haplotype diversity. A comparison the recombination map of chromosome 25 (D06) also reveals a suppression of recombination in this region for the intraspecific populations, but not in some of the interspecific PS7xNemX RIL and TM-1x3-79 RIL populations (**Figure 2.10**).This suggests that interspecific germplasm can be used to introduce new genetic material at this location and perhaps concomitantly alter the recombinational landscape. At this point, it is not known if such differences in recombinational profiles are heritable, how so, and if effects would be cis- or trans-acting. Answers, however, should be sought through future recombination and introgression research.

CHAPTER III

CIS-FEATURES ASSOCIATED WITH MEIOTIC RECOMBINATION IN *GOSSYPIUM* GENUS

Introduction

A barrier to germplasm introgression and other breeding efforts arises when the ability to generate novel combinations alleles of phenotypic importance is constrained, as happens when meiotic recombination is lacking or exceptionally rare between desirable alleles in repulsion, and/or a failure to transmit and recover the desired recombinants. Low rates of recombination around introgressed favorable alleles can prevent their recovery in favorable combinations of alleles of nearby loci. Decreased rates of crossovers are generally seen in heterochromatin and centromeric regions but are highly variable across the chromosome with certain exceptions such as at the boundaries that demarcate haplotype blocks [89]. Regions that recombine at elevated rates relative to the genome-wide rates are termed *recombination hot spots*, while those that recombine at low rates relative to the genome wide rate are termed *recombination cold spots*. These two recombination categories, or subtypes, are important for understanding genetic variation and how genomes evolve as well as by determining the efficacy by which recombination can segregate deleterious alleles.

Several *cis* features can be associated with recombination rates such as the density of transposable elements and their epigenetic regulation [47, 90]. There is a negative correlation between TE density and recombination, but the strength of the correlation can vary depending the type of TE [47]. The accumulation of TEs is also greater at recombination cold spots and is believed to provide a positive feedback loop by further spreading the suppression of recombination through epigenetic regulation. The epigenetic regulation of TEs can be achieved through DNA methylation [90], RNA silencing [91], and histone modification [43, 92]. This epigenetic regulation reduces the numbers of ectopic insertions of TEs and promotes genome stability.

Recombination frequencies can also modulate the efficacy of selection pressures on advantageous and deleterious mutations. Disadvantageous alleles that arise through mutations can be more differentially and thus more effectively purged from the genome through recombination-empowered segregation. Areas under low recombination can lead to co-inheritance of deleterious alleles with positively selected neighboring allele, which may increase the genetic load of the genome across generations. Deleterious mutations are generally associated with nonsynonymous substitutions as they result in a change of the amino acid sequence of a protein or by the creation of premature stop codons. Synonymous mutations on the other hand do not alter the amino acid sequence, but can result in a phenotypic change by altering the splicing pattern [93] of an mRNA transcript through mutating the intron-exon junction site or by altering the codon usage bias [94, 95].

Recombination rates have also been demonstrated to be influenced by environmental stimuli [96, 97] and in *G. hirsutum*, a correlation between recombination rates and functional gene categorizations has been made [98]. The genes and biological pathways associated with homologous recombination hot spots and cold spots between species can be used to assess the level of genetic diversity in regard to gene content at these specific regions. Additionally, inversions, if heterozygous, might greatly exacerbate low rates by sabotaging the viability of single crossover products, and segments with low-crossover rates are less likely to form double or higher crossover events. It should be noted that recombination in this study is estimated among progeny, not directly in meiotic products, so it is possible that effects on recombination may be skewed by factors affecting the transmission and viability of single-crossover products, and those factors could differentially affect the major chromosome/chromatin regions. Understanding the distribution of genomic *cis* acting features such as TE density, gene content, and nucleotide substitutions within recombination cold spots and hot spots may facilitate in eliciting mechanisms that can be used to modulate the recombination landscape of cotton.

Materials and Methods

Plant Material and Genotyping

Two interspecific mapping populations consisting of 59 *G. hirsutum* (TM-1) x *G. mustelinum* BC_1F_1 and 85 *G. hirsutum* (TM-1) x *G. tomentosum* BC_1F_1 where grown over the summer of 2016 in College Station, Tx. Young leaf tissue was collected from these populations and DNA was isolated using a Machery-Nagel Plant Nucleo-spin kit (Pennsylvania) according to the manufacturer's instructions. Isolated DNA was quantified using PicoGreen® and then diluted to 50ng/µl. Genotyping was conducted using the CottonSNP63K Array at Texas A&M University [65] and SNPs were called using Illumina's® GenomeStudio (v2.0) software resulting in 63,058 SNPs functional and non-functional SNPs per population. Genotyping data from a previously published consensus map [99] was utilized as well.

Development and Mapping of Intraspecific and Interspecific Linkage Maps

Genotyping data were processed for quality control by removing SNPs that had more than 10 percent missing data or had a minor allele frequency (MAF) greater than 5 percent. Individuals with missing data greater than 10 percent were also removed. Genotype data were converted to ABH format using a panel of parental genotypes [65] where "A" represents an allele descendant of parent 1, "B" an allele descendent of parent 2, and "H" represents a heterozygous genotype.

Processed genotype data were imported into JoinMap 4.1 [66] and linkage groups were coalesced using the independence logarithm of the odds (LOD) function within the grouping tree tool. Grouping processes was reiterated with increasing LOD thresholds until 26 coherent linkage groups were established. SNPs that did were not integrated into a linkage group were removed. The maximum likelihood algorithm with a maximum recombination frequency of 50% and the Kosambi function were used to order the SNPs within each of the linkage groups. SNPs with a nearest neighbor (NNS) stress greater than 5 centiMorgans (cM) were removed and reordering of markers was conducted until markers exhibited a NNS less than 5. Array ID sequences retained in the final linkage maps where aligned to both their respective species' sequence assemblies were applicable. The *G. hirsutum* x *G. barbadense* F_2 array IDs were mapped to both the JGI *G. hirsutum* v2.1 and JGI *G. barbadense* v1.0 sequence assemblies; the

G. hirsutum x *G. tomentosum* BC_1F_1 array IDs were aligned to the both the JGI *G. hirsutum* v2.1 and JGI *G. tomentosum* sequence assembles; and the *G. hirsutum* x *G. mustelinum* BC_1F_1 array IDs were aligned to the JGI *G. hirsutum* v2.1 and *G. mustelinum* v1.0 sequence assemblies. The intraspecific consensus map array IDs were only mapped to the JGI *G. hirsutum* v2.1 sequence assembly. All alignments were conducted using BLASTn (v2.7.1+) [68] with a minimum e-value cutoff of 1e-10 and with the dust and soft masking parameters disabled. Array IDs that aligned to both a homeolog relative to the linkage map and to the linkage map homolog were ultimately mapped to the homolog i.e. the linkage map was the deciding factor in ambiguously aligned markers. Array IDs that mapped to the same location were filtered based on BIT score.

Recombination Rate Estimation and Categorization

The recombination rates of the three linkage mapping populations were estimated by the SW method as previously described in chapter 2. The SW regression method was used as opposed to the LOESS regression to remove the occurrence of false positives that may be generated by the interpolation of recombination rates within areas of the genome that do not provide sufficient SNP density to accurately estimate the recombination rate. Interpolation of false recombination at regions under increased linkage may occur if neighboring SNPs within close proximity are recombining in order to smooth the fit of the regression [100]. SNPs were then paired into non-overlapping groups and assigned the recombination rate of the first SNP in the pair. This effectively translated the cM distance between neighboring pairs of SNPs into a recombination rate as estimated by the SW regression. Recombination rates were then filtered for complete linkage (*D'* = 1) in order to not bias the distribution of recombination rates by haplotype blocks; this procedure effectively removed all SNPs, except one, with identical cM positions. The distribution of the recombination rates was determined using R and then pairs of SNPs with recombination rates above the 97.5 percentile were categorized as being highly recombinant (hereafter referred to as recombination hotspots) and pairs of SNPs with recombination rates below the 2.5 percentile were categorized as being lowly recombinant (henceforth referred to as recombination coldspots) in order to capture a total of 5% of the greatest recombination rate variation within the genome.

Ascertainment and Generation of Homologs Sequence Coordinates

The genomic coordinates of the recombination cold spots and hot spots were converted into BED files which contained the chromosome, start, and stop position (bp) of the SNP pairs. Array IDs from the 3 interspecific linkage maps that were successfully aligned in both parental species were paired according to array ID. Homologs SNPs were then used to generate BED files that were homologous to the previously generated BED files (**Figure 3.1**). Ultimately, each interspecific mapping population resulted in the generation of a pair of homologous BED files: 1 for *G. hirsutum* coordinates (common parent) and a 2nd for its respective parental species. Each BED file was further divided into recombination cold spot coordinates and recombination hot spot coordinates. Only one BED file was generated for the intraspecific consensus map, also divided into hot and cold recombination spots.

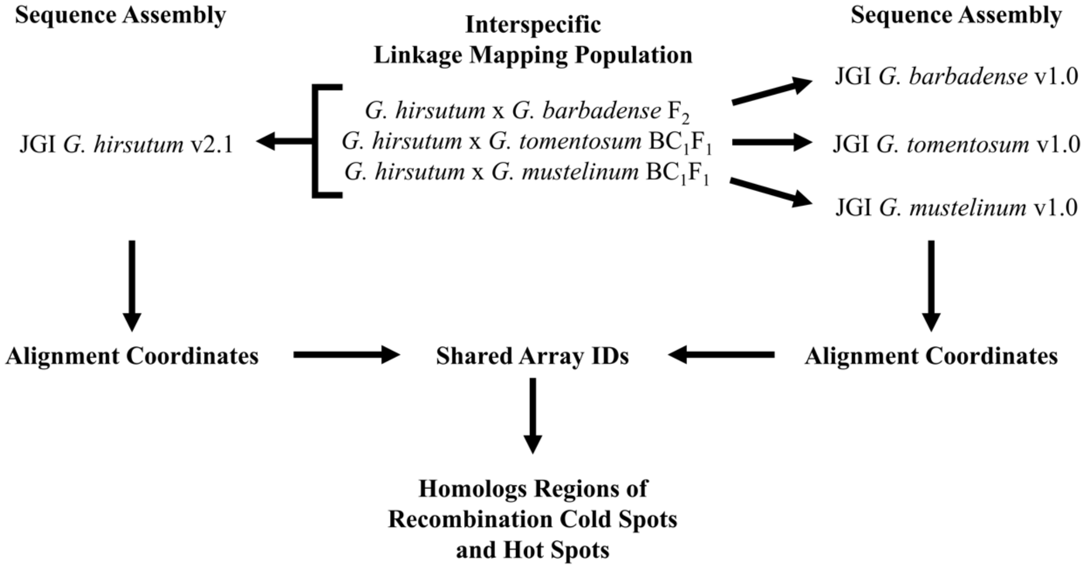

Figure 3.1 Flow chart indicating the alignment process of shared array IDs to the various JGI sequence assemblies.

Relationship of Homologous Sequences

The length of the DNA sequence within each flanking pair of SNPs was measured using the genome coordinates of the respective BED file. The sequence lengths were then compared to the sequence length of the respective homologs sequence using R to measure the differences in sequence lengths between homologous sequences for both recombination cold spots and recombination hot spots. A non-parametric Wilcoxon rank sum test was conducted on the differences in sequence lengths between recombination cold spots and hot spots following a Shapiro's test for normality. A test for association between homologs sequences lengths for both recombination cold spots and hot spots was conducted using Pearson's correlation coefficient (r^2) using R.

Gene Enrichment Analysis

Global gene lists for *G. hirsutum*, *G. barbadense*, *G. tomentosum*, and *G. mustelinum* were retrieved from their respective annotated sequence assemblies. The specific set of genes contained within recombination cold spots and hotpots for all mapping populations were retrieved by intersecting each respective BED file with its respective sequence assembly annotated general feature file (GFF3) using bedtools intersect (v2.26.0) which retrieves overlaps between two files [101]. Each GFF3 file was then filtered to contain genes only. Gene ontology (GO) terms for each sequence assembly were retrieved from text files containing the annotated gene functions, gene IDs, and their respective GO terms.

The global gene list and its respective subset of recombination cold spots and hot spots gene list were then used to conduct gene enrichment analysis using the R package topGO (v2.37.0) [102] which takes into account the GO term hierarchy when performing gene enrichment which theoretically results in fewer false positives [103, 104]. Genes enrichment for biological pathways was conducted by using topGO's "weigt01" algorithm which traverses the GO hierarchy from bottom to top in order to de-emphasize GO terms higher in the hierarchy while simultaneously comparing the significance scores of parent and child nodes. A Fisher's exact test was then performed on gene counts followed by correction for the false discovery rate (FDR) [105].

Transposable Element Densities

General feature files containing annotated repetitive sequences were retrieved from *G. hirsutum*, *G. barbadense*, *G. tomentosum*, and *G. mustelinum* sequence assemblies. The GFF3 files were intersected against themselves e.g. *G. barbadense* GFF3 file vs *G. barbadense* GFF3 file, using bedtools intersect (v2.26.0) while removing overlaps in respect to strandedness e.g. removes duplicate annotation on opposite strand if applicable. The processed GFF3 files where then filtered for six major classes of transposable elements (TE): long interspersed nucleotide elements (LINEs), short interspersed nucleotide elements (SINEs), long terminal repeats (LTRs), rolling circle or helitrons (REs), mutator-like transposable elements (MULEs), and Harbingers also known as P instability factor-like (PIF) transposons. The filtered GFF3 files where then intersected with their respective BED file in order to retrieve the TEs within the specified regions. The length (number of bases) of each TE within each region was summed and divided by the total length of the sequence from which the TEs where retrieved from in order to calculate the TE density within that region.

Processed GFF3 TE files containing genome wide TE annotations from each sequence assembly were divided into non-overlapping 500 Kb regions and TE densities within each 500 Kb windows were calculated using a custom script. These TE densities were then used as controls. A non-parametric Dunn test was conducted between recombination specific TE densities and genome wide TE densities following a Shapiro's test for normality.

Ortholog Copy Number and Synteny

Three files containing the orthologs between *G. barbadense* and *G. hirsutum*, *G. tomentosum* and *G. hirsutum*, and *G. mustelinum* and *G. hirsutum* were retrieved from their respective JGI genome assembly annotations. The ortholog lists were filtered for genes within recombination cold spots and hot spots using the previously generated gene enrichment lists. Analysis of genome-wide ortholog copy number was conducted similarly. A non-parametric Dunn test [106] was conducted between recombination-specific ortholog copy number and the genome-wide copy number following an Anderson-Darling test for normality [107]. Ortholog copy number was then categorized into copy-number type (here forth referred to as ortholog depth) for further comparison.

The primary nucleotide coding sequence FASTA and GFF3 genomic coordinate files were retrieved from the *G. hirsutum, G. barbadense*, *G. tomentosum*, and *G. mustelinum* sequence assembly annotations. Files were then filtered to for genes and gene coordinates that corresponded to recombination cold spots and hots. Orthologs within recombination cold spots and hotspots were then aligned to one another using the software MCSCan [108] utilizing the filtered primary coding sequences and genomic coordinates (GFF3 files). Orthologous pairs were defined using the previously retrieved ortholog files in order to ensure correct pairing of orthologs.

Results

Recombination Landscape

The 1-Mb non-overlapping sliding window with a threshold of 4 SNPs for non-linear regression was used to estimate the genome-wide recombination rates of one intraspecific consensus map and three interspecific populations (**Figure 3.2-3.5**). The recombination landscape was found to be nonuniform across individual chromosomes with a majority of recombination occurring within the sub-telomeric ends of the chromosome and decreasing towards the pericentromeric and centromeric regions. The interspecific populations exhibited an increase in recombination within the pericentromeric regions of chromosomes within the D subgenome relative to the A subgenome (**Figure 3.3, Figure 3.4,** & **Figure 3.5**). Furthermore, the A-D subgenome differences in size and gene density are accentuated in the pericentromeric regions. The highest recombination rate was seen in the *G. hirsutum* x *G. barbadense* F_2 mapping population on chromosome D05 with a rate of 49.79 cM/Mb and the lowest recombination rate was seen in the *G. hirsutum* x *G. tomentosum* BC_1F_1 mapping population on chromosome A01 with a rate of 0.01 cM/Mb (**Figure 3.3** & **Figure 3.5**).

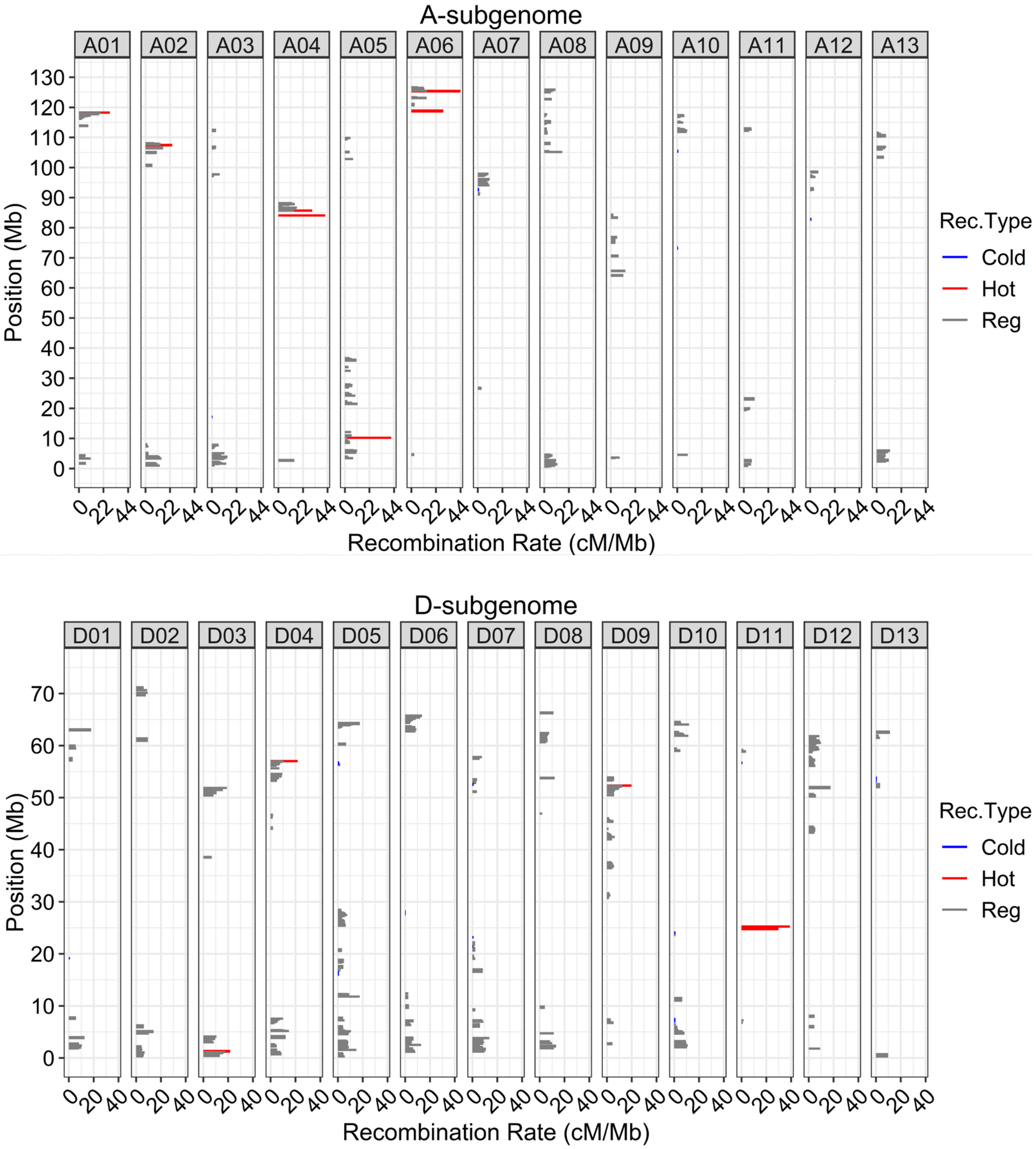

Figure 3.2 Genome-wide recombination map of the intraspecific consensus mapping population. Local recombination rates where estimated using a non-overlapping 1-Mb sliding window. Recombination hot spots are shown in red and recombination cold spots are shown in blue. Regular recombination rates are shown in grey.

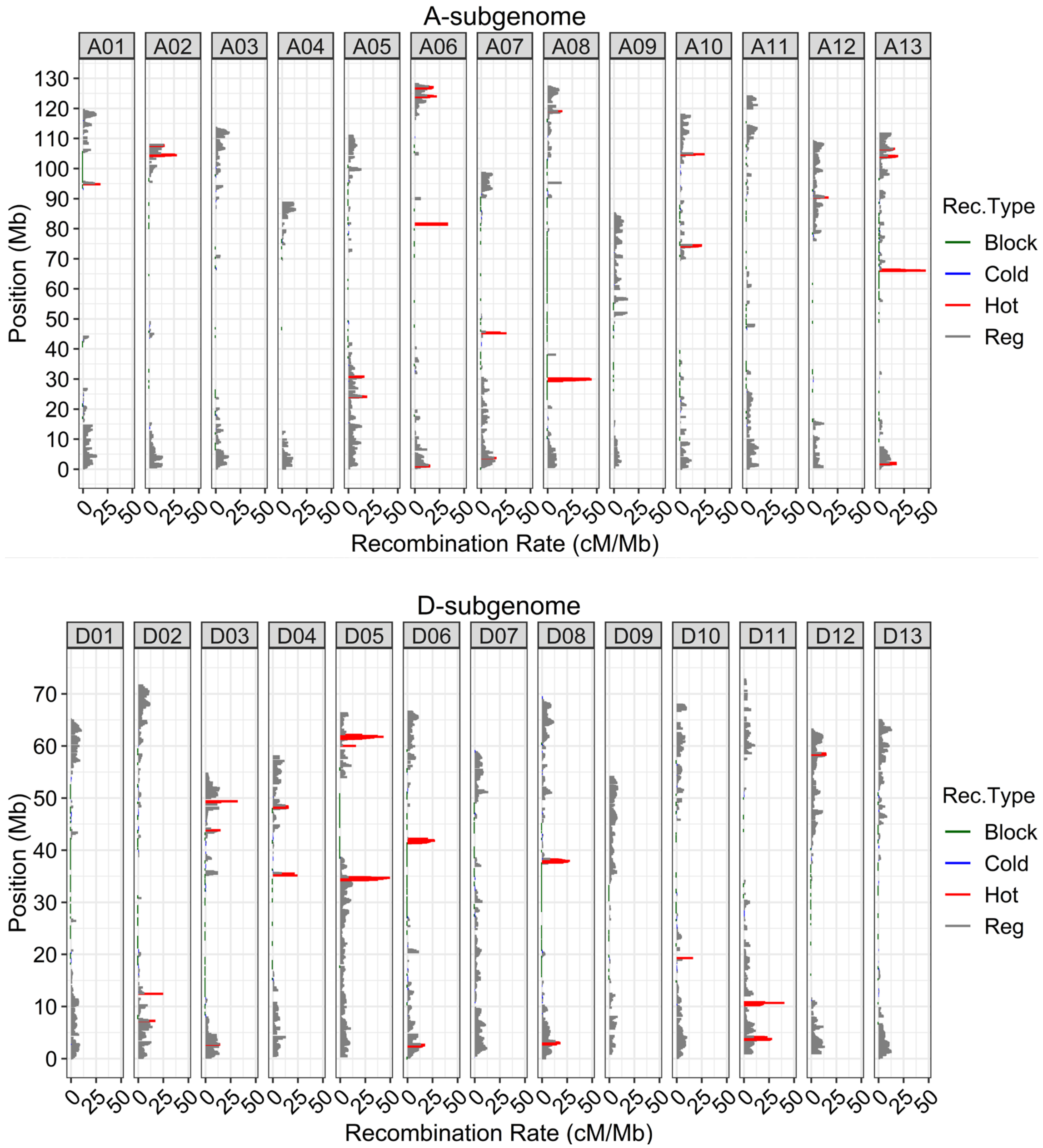

Figure 3.3 Genome-wide recombination map of the *G. hirsutum* x *G. barbadense* F_2 population. Local recombination rates where estimated using a non-overlapping 1-Mb sliding window. Recombination hot spots are shown in red, recombination cold spots are shown in blue, and linkage blocks (*D'* = 1). Regular recombination rates are shown in grey.

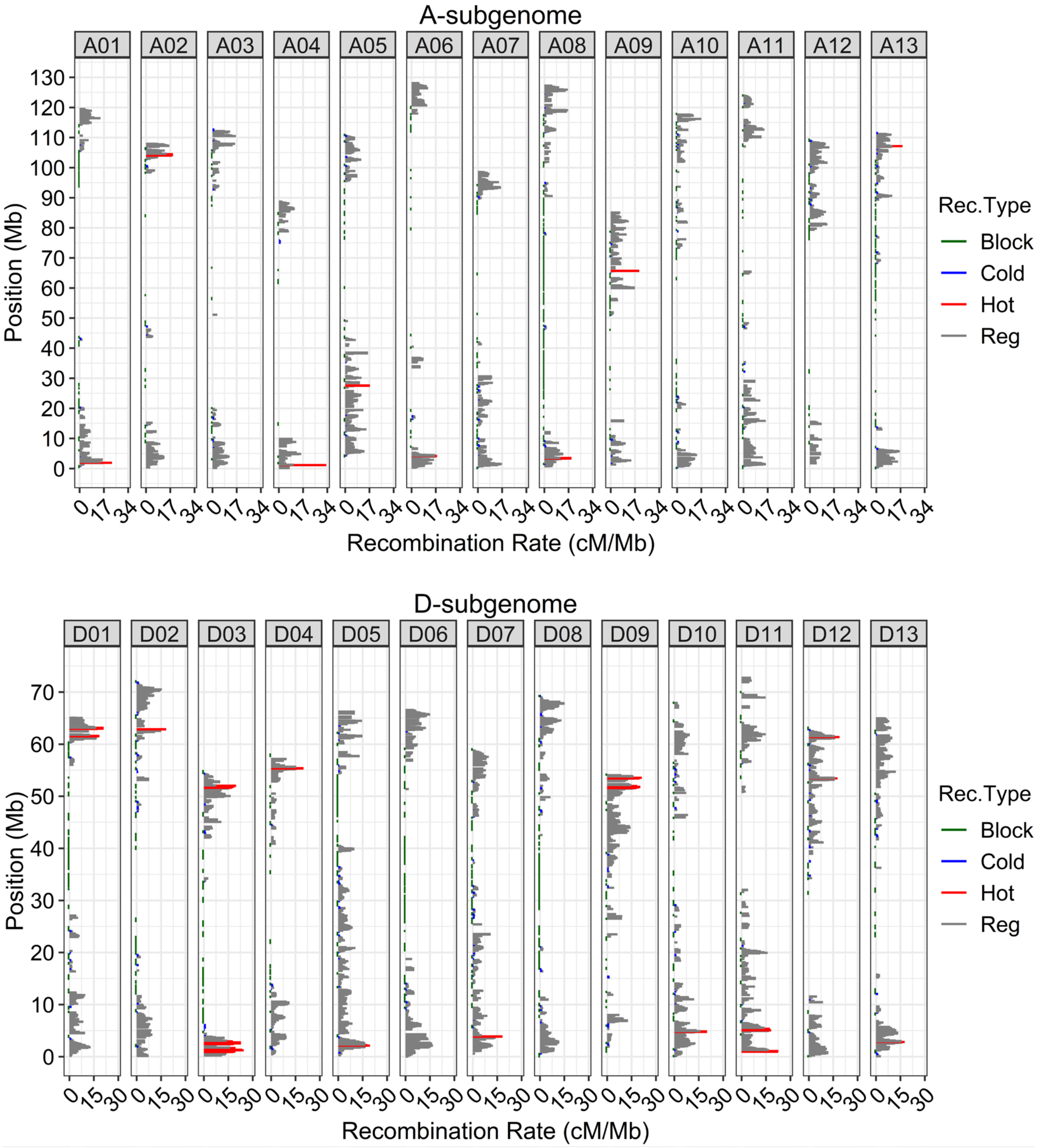

Figure 3.4 Genome-wide recombination map of the *G. hirsutum* x *G. tomentosum* BC_1F_1 population. Local recombination rates where estimated using a non-overlapping 1-Mb sliding window. Recombination hot spots are shown in red, recombination cold spots are shown in blue, and linkage blocks (*D'* = 1). Regular recombination rates are shown in grey.

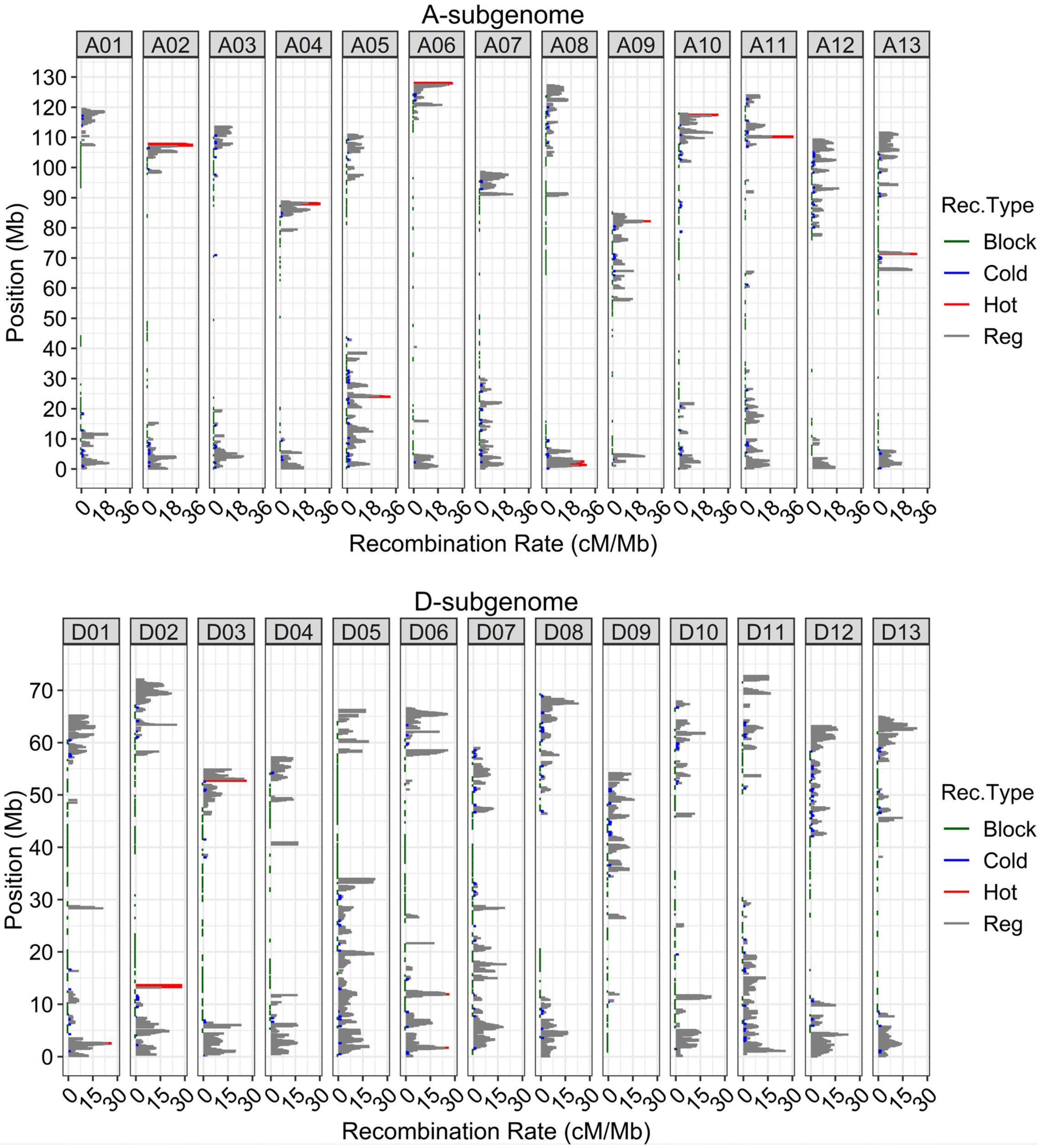

Figure 3.5 Genome-wide recombination map of the *G. hirsutum* x *G. mustelinum* BC$_1$F$_1$ population. Local recombination rates where estimated using a non-overlapping 1-Mb sliding window. Recombination hot spots are shown in red, recombination cold spots are shown in blue, and linkage blocks (*D'* = 1). Regular recombination rates are shown in grey.

Categorizing Recombination Cold Spots and Hot Spots

Alignment of marker array IDs to their respective parent 2 (**Figure 3.1**) for pairs of SNPs categorized as recombination cold spots and hot spots (**Figure 3.6**) were paired to their parent 1 homologs. Not every pair of array IDs was successfully aligned to its respective homolog (**Table 3.1**) and ultimately only the array IDs that were successfully mapped in both parents were used for further downstream analysis. Homologous sequence lengths of array IDs mapped in both parents were compared to one another to assess homologous relationship based on whether the sequence lengths were significantly different in size. All sequence lengths had a Pearson's coefficient of correlation greater than 95% and there was no statistically significant difference (p-value > 0.05) in the difference in sequence length between recombination cold spots and recombination hot spots (**Figure 3.7**). This suggests that there are no major insertion or deletion events associated with regions undergoing any rate of recombination.

A.

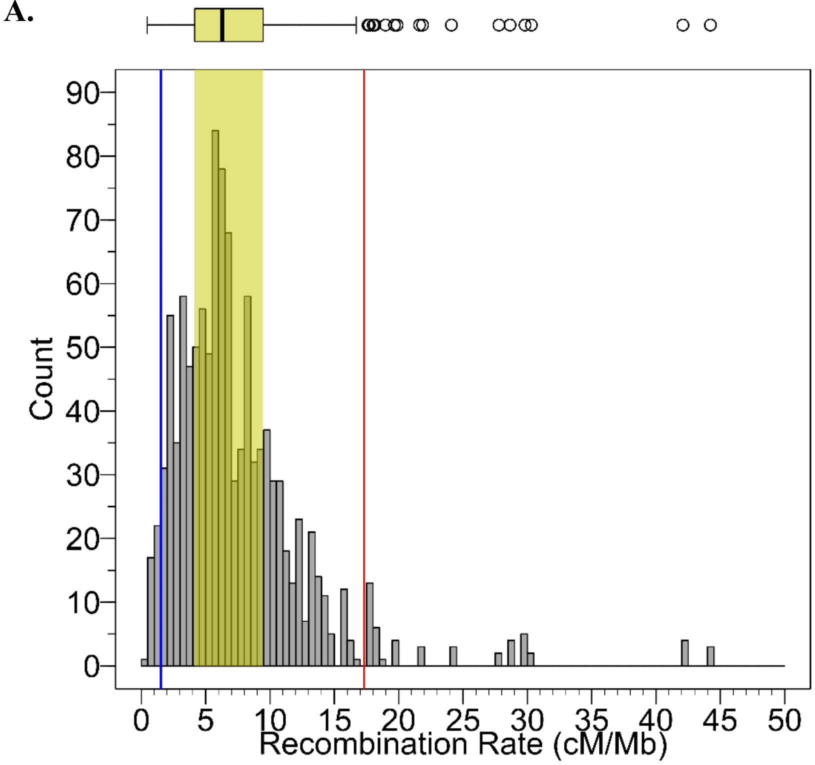

Figure 3.6 Recombination rate distributions for four mapping populations. A) Intraspecific consensus map and interspecific maps **B)** GhxGb F_2, **C)** GhxGt BC_1F_1, and **D)** GhxGm BC_1F_1. Rates above than the 97.5 percentile (red line) are categorized as recombination hotspots, whereas rates below the 2.5 percentile (blue line) are categorized as recombination coldspots.

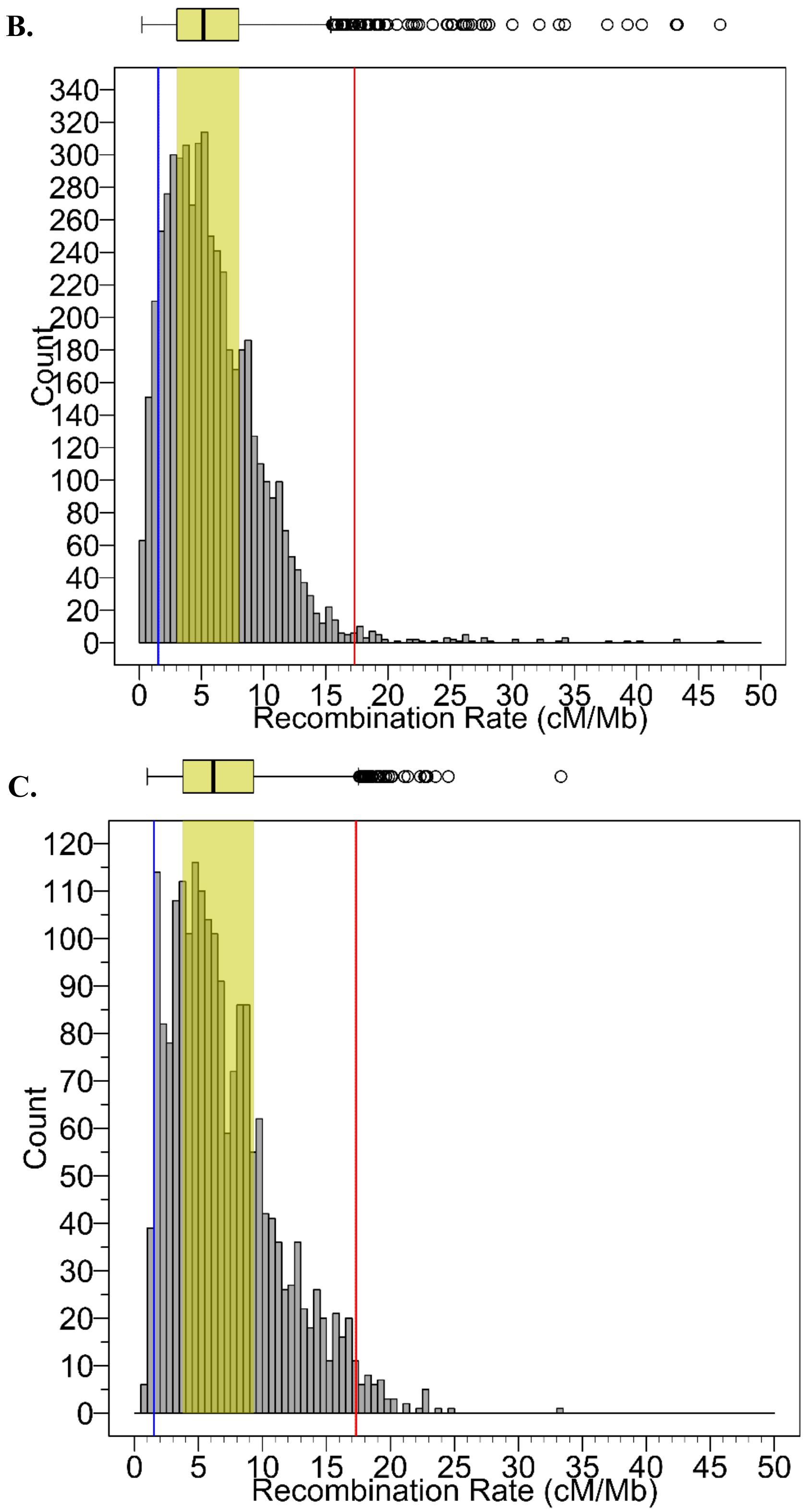

Figure 3.6 (continued)

D.

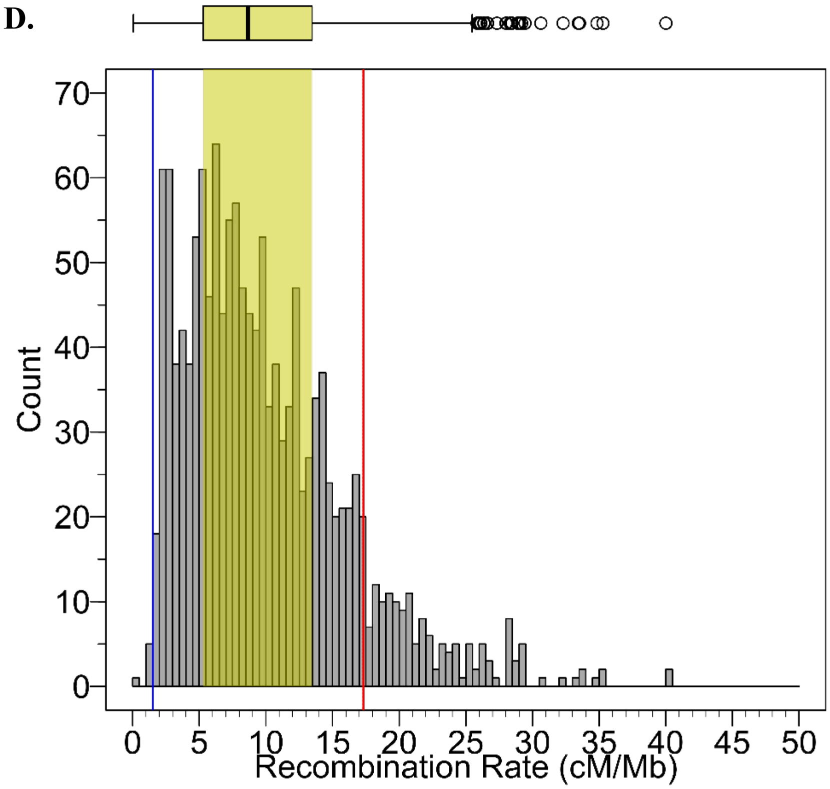

Figure 3.6 (continued)

Table 3.1 Numbers of recombination cold spots and hot spots successfully aligned to the genome assemblies of parental species for three interspecific maps and one consensus map.

	Recombination Type	ConMap	GhxGb	GhxGt	GhxGm
Mapped in Parent 1	Cold	28	127	49	36
	Hot	28	127	49	36
Mapped in Parent 2	Cold	NA	123	48	33
	Hot	NA	117	49	36

A.

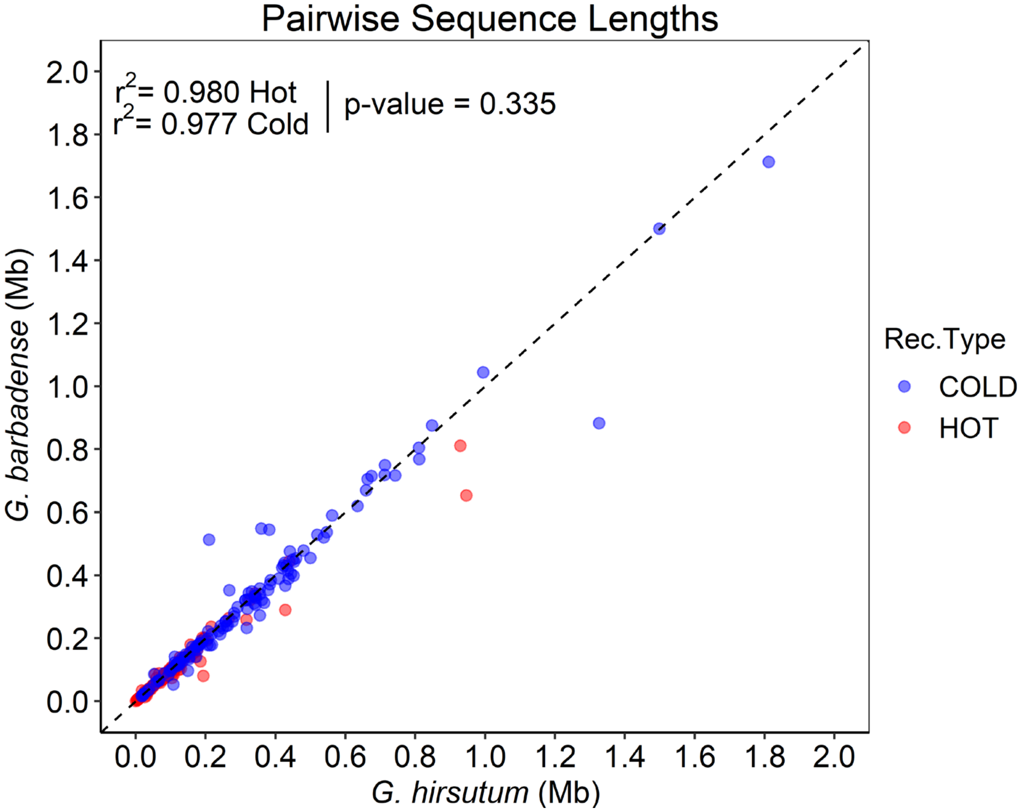

Figure 3.7 Sequence length comparisons of recombination regions between parent species. Sequence length (Mb) for *G. hirsutum* is shown on the x-axis while the homologous sequence for **A)** *G. barbadense*, **B)** *G. tomentosum*, and **C)** *G. mustelinum* are shown on the y-axis of each respective graph. Sequence lengths comprising recombination cold spots are shown in blue, whereas sequence lengths comprising recombination hot spots are shown in red. Pearson's coefficient of correlation for each recombination type is shown at the top-left of each graph.

B.

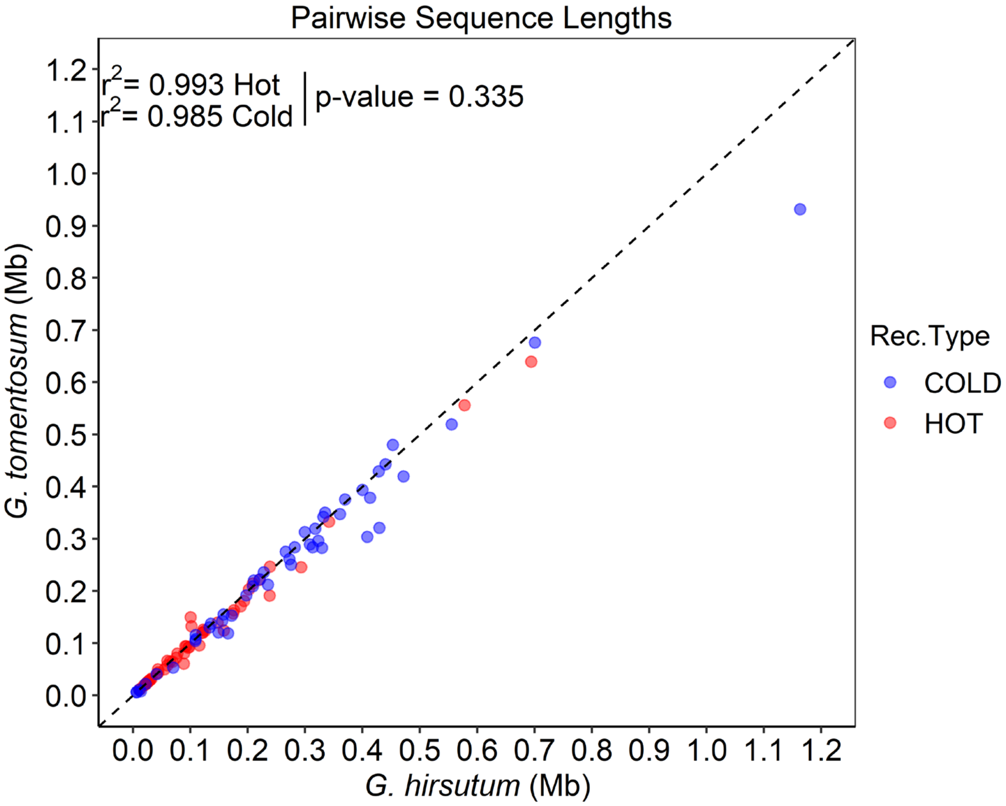

Figure 3.7 (continued)

C.

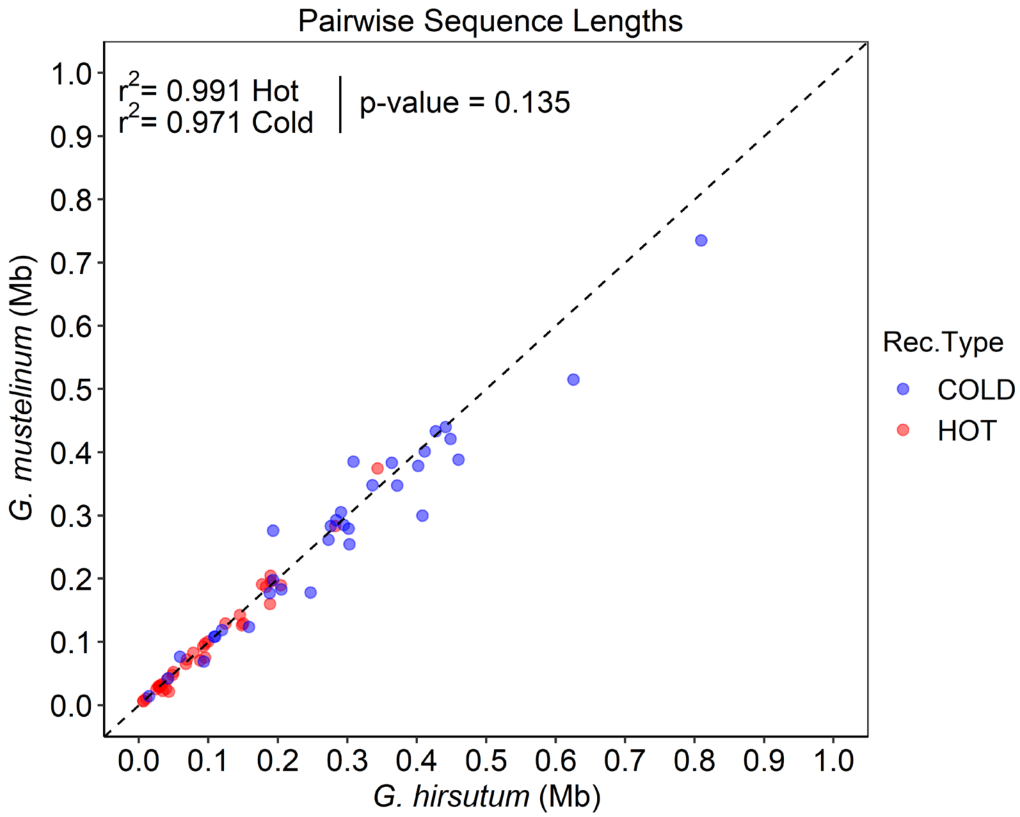

Figure 3.7 (continued)

Gene Enrichment Analysis of Recombination Cold Spots and Hot Spots

Analysis of genes located within regions recombining at a low or high rate revealed biological pathways associated with these regions. It has been previously demonstrated [98] that regions correlated with high recombination are associated with metabolic processes involved in response to abiotic and biotic stimuli, while regions correlated with low recombination are associated with mitotic and meiotic processes. The inferred explanation was that recombination is required to adapt to a changing environment while simultaneously maintaining cell cycle stability. This previously detected relationship between recombination type and biological pathway was not detected based gene enrichment analysis of the cotton consensus map. Gene enrichment analysis of the intraspecific consensus map resulted in only four biological pathways being significantly associated (q-value < 0.05) with recombination hot spots following multiple test correction (**Figure 3.8)**; there were no clear distinctions between recombination subtypes or among biological pathways, even when not correcting for multiple tests (p-values < 0.05). This lack of detection may be due to the fact that the consensus map was constructed using three intraspecific populations, which may have biased the distribution of recombination cold spots and hot spots.

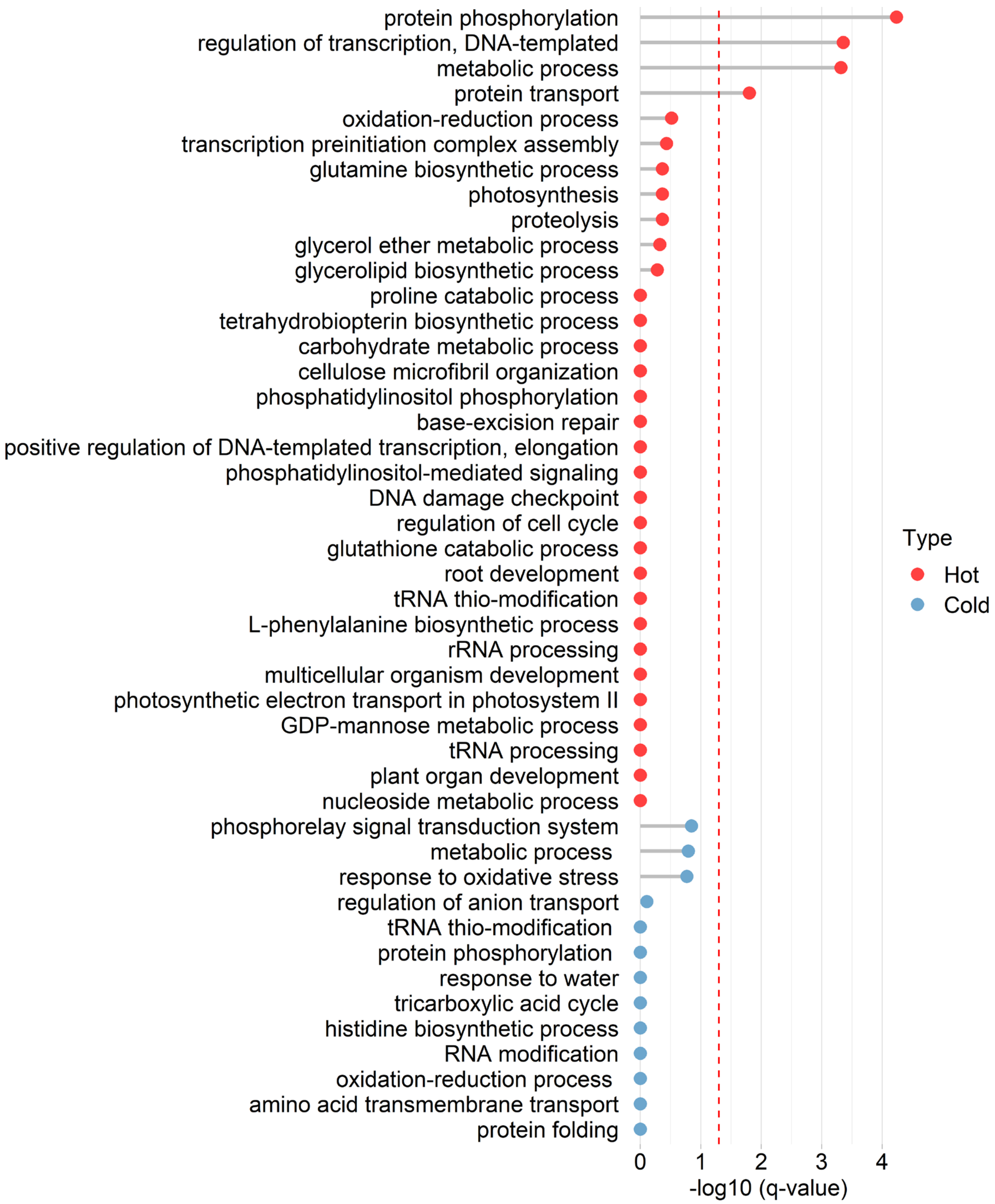

Figure 3.8 Gene enrichment for recombination types for the intraspecific consensus mapping population. Recombination hot spots are shown in red while recombination coldspots are shown in blue. All biological pathways shown have a p-value of less than 5% and biological pathways with a q-value (-log10) of less than 5% are represented to the right of the red dashed line.

The gene enrichment analysis of the *G. hirsutum* x *G. barbadense* F_2 population is congruent with the claim that recombination hot spots that are correlated with biological pathways involved in environmental stimuli, while recombination cold spots are correlated with pathways involved with the cell cycle (**Figure 3.9**). Gene enrichment in *G. hirsutum* reveals that recognition of pollen is significantly associated (q-value < 0.05) with recombination hot spots following multiple test correction. Prior to the correction for multiple tests, terpenoid biosynthetic process, defense response to bacterium, and regulation of systemic acquired resistance biological pathways are associated with hot spots (p-value < 0.05) (**Figure3.9A**). Gene enrichment analyses of the homologous regions in *G. barbadense* reveals that the terpenoid biosynthetic pathway is significantly associated (q-value < 0.05) with recombination hot spots following multiple test correction while regulation of systemic acquired resistance, photosynthesis and light harvesting, response to stress, and defense response to bacterium are associated (p-value < 0.05) not following multiple test correction (**Figure3.9B**). Terpenoids are of specific interest to cotton breeders as they are commonly found in cotton, e.g., the terpenoid aldehyde gossypol, which acts as a pesticide to insects, and have been shown to accumulate following mechanical damage [109]. Although not statistically significant following multiple test correction, cell cycle, nucleotide-excision repair, cell cycle arrest, regulation of DNA replication, cell growth, and regulation mitotic chromosome condensation were biological pathways associated (p-value < 0.05) with recombination cold spots in either *G. hirsutum* or *G. barbadense* (**Figure3.9 A-B**).

A.

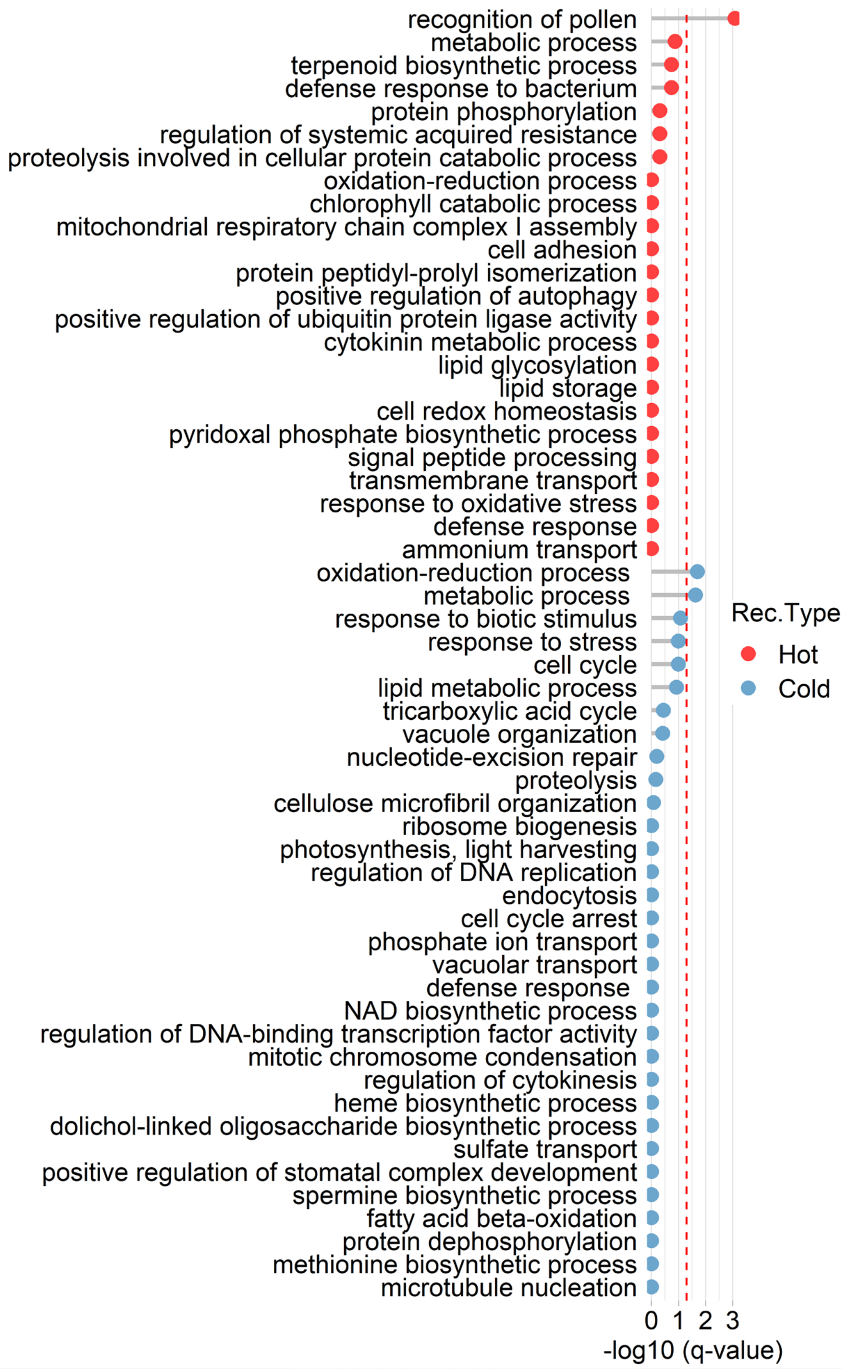

Figure 3.9 Gene enrichment results for the *G. hirsutum* x *G. barbadense* F$_2$ mapping population. A) Genes were taken from the JGI *G. hirsutum* sequence assembly and from the homologous sequence regions from the **B)** JGI *G. barbadense* sequence assembly were used for gene enrichment. Recombination hot spots are shown in red while recombination coldspots are shown in blue. All biological pathways shown have a p-value of less than 5% and biological pathways with a q-value (-log$_{10}$) of less than 5% are represented to the right of the red dashed line.

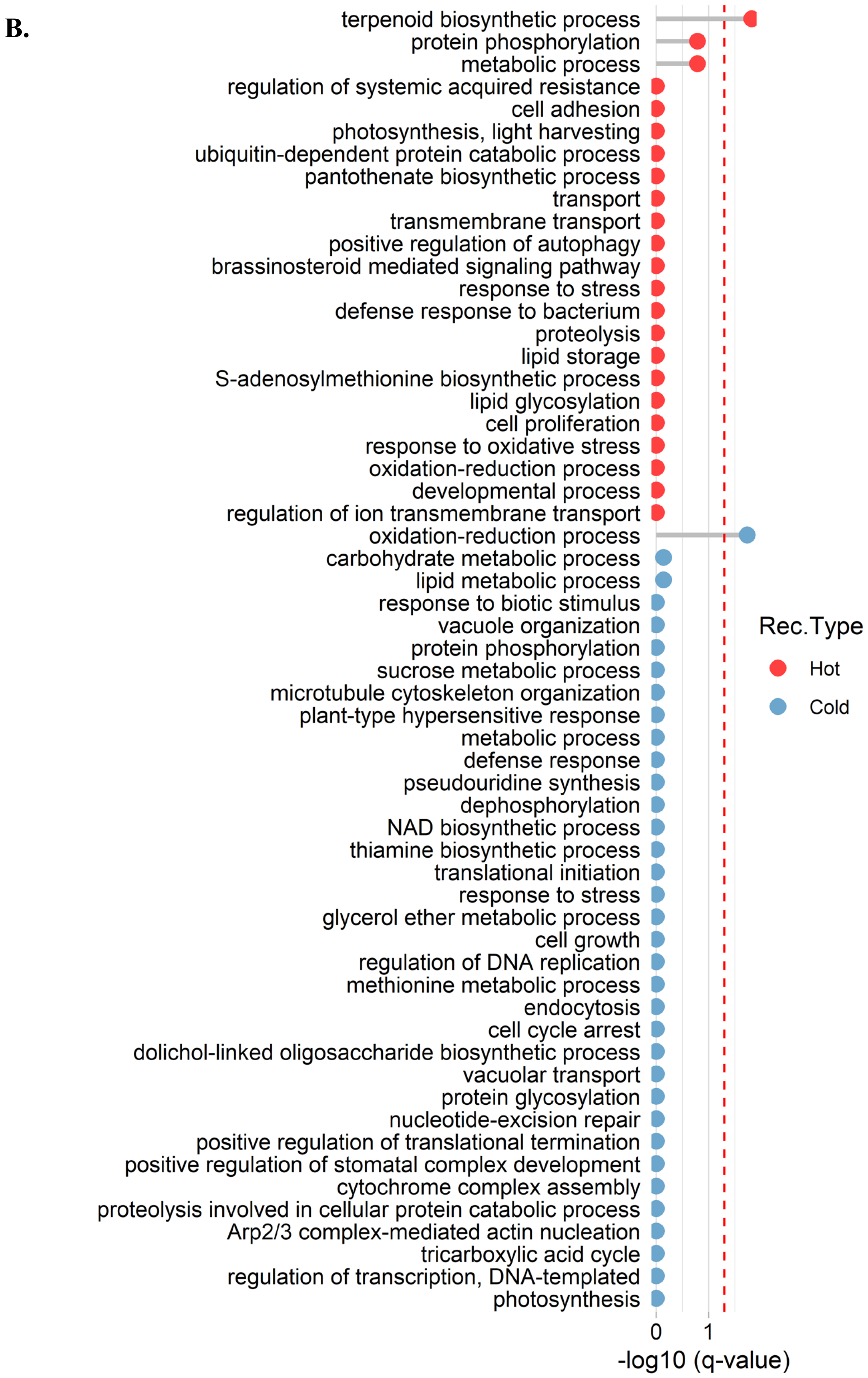

Figure 3.9 (continued)

Gene enrichment analysis of the *G. hirsutum* x *G. tomentosum* BC_1F_1 population revealed two biological processes associated with recombination hot spots (q-value < 0.01) following multiple test correction with no explicit pathways correlated with environmental stimuli (**Figure 3.10B**). Although not statistically significant following multiple test correction, biological pathways involved in the cell cycle such as reciprocal meiotic recombination, mismatch repair, and double-strand break repair via homologous recombination were correlated (p-value < 0.05) with recombination cold spots (**Figure3.10AB**). Only one biological pathway was found to be significantly associated with recombination following multiple test correction in the *G. hirsutum* x *G. mustelinum* BC_1F_1 (**Figure 3.11B**) with no discernible correlations between recombination type and environmental stimuli or cell cycle pathways.

A.

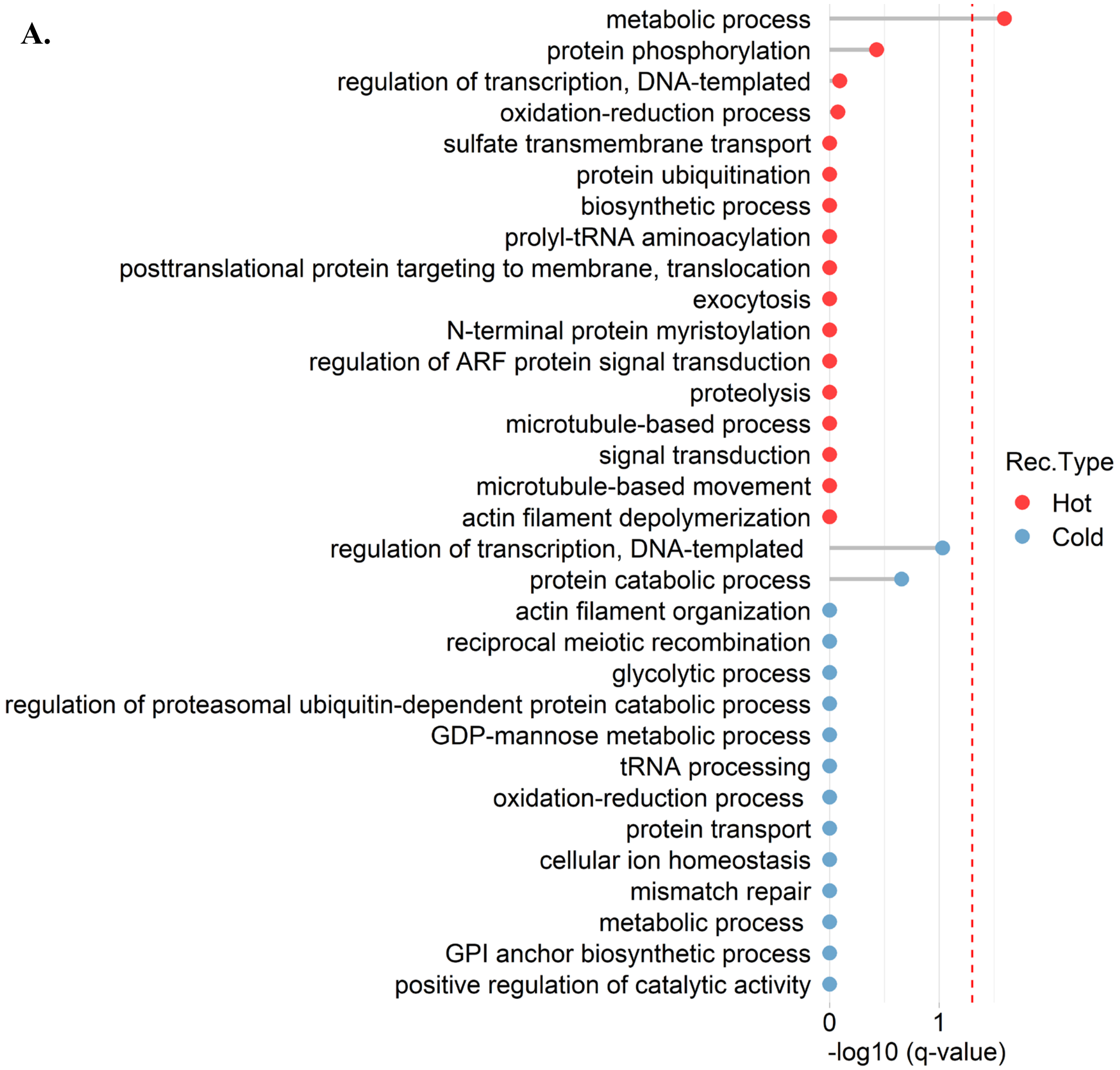

Figure 3.10 Gene enrichment results for the *G. hirsutum* x *G. tomentosum* BC_1F_1 population. A) Genes were taken from the JGI *G. hirsutum* sequence assembly and from the homologous sequence regions from the **B)** JGI *G. tomentosum* sequence assembly were used for gene enrichment. Recombination hot spots are shown in red while recombination coldspots are shown in blue. All biological pathways shown have a p-value of less than 5% and biological pathways with a q-value ($-log_{10}$) of less than 5% are represented to the right of the red dashed line.

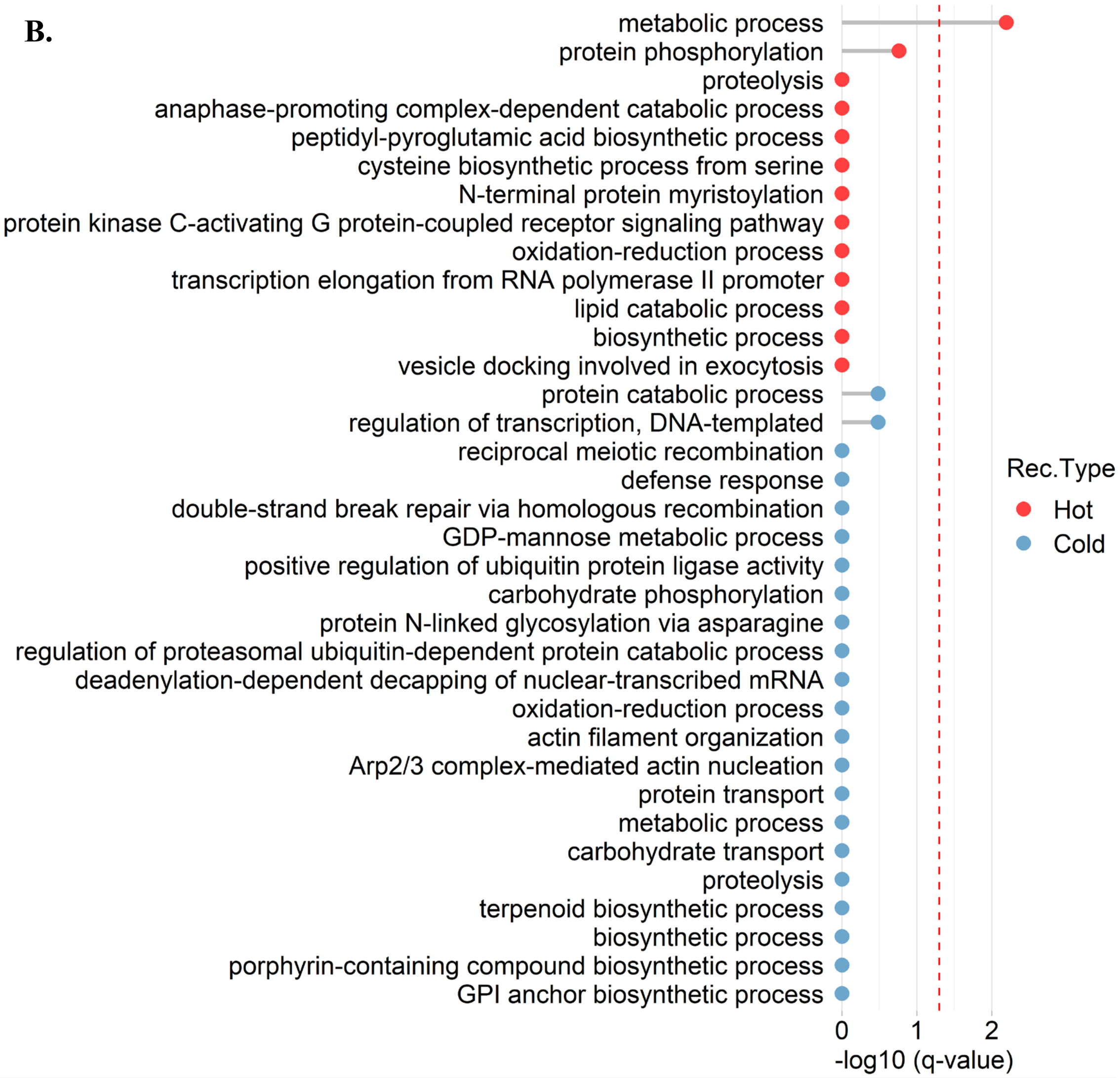

Figure 3.10 (continued)

A.

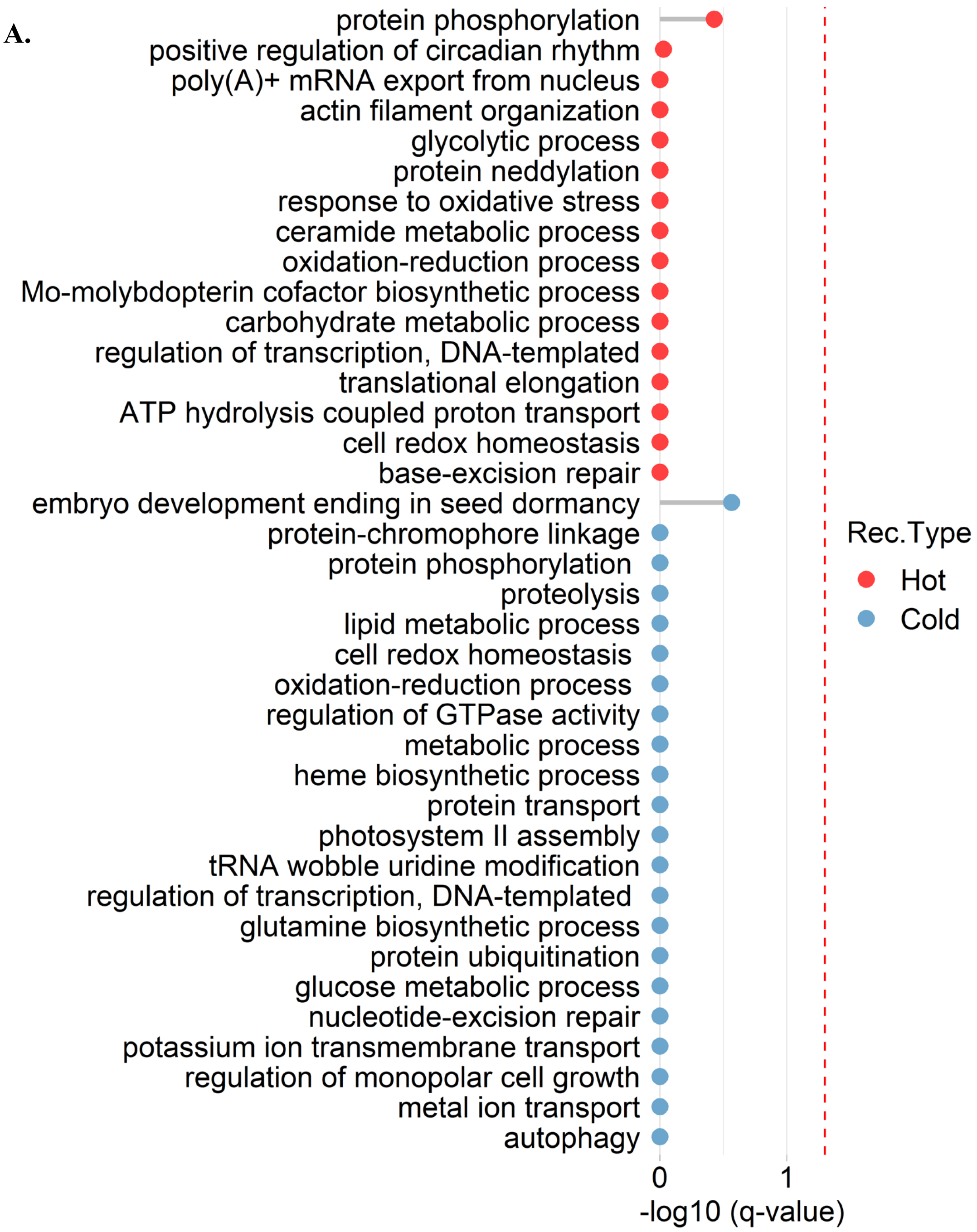

Figure 3.11 Gene enrichment results for the *G. hirsutum* x *G. mustelinum* BC$_1$F$_1$ population. A) Genes were taken from the JGI *G. hirsutum* sequence assembly and from the homologous sequence regions from the **B)** JGI *G. mustelinum* sequence assembly were used for gene enrichment. Recombination hot spots are shown in red while recombination coldspots are shown in blue. All biological pathways shown have a p-value of less than 5% and biological pathways with a q-value (-log$_{10}$) of less than 5% are represented to the right of the red dashed line.

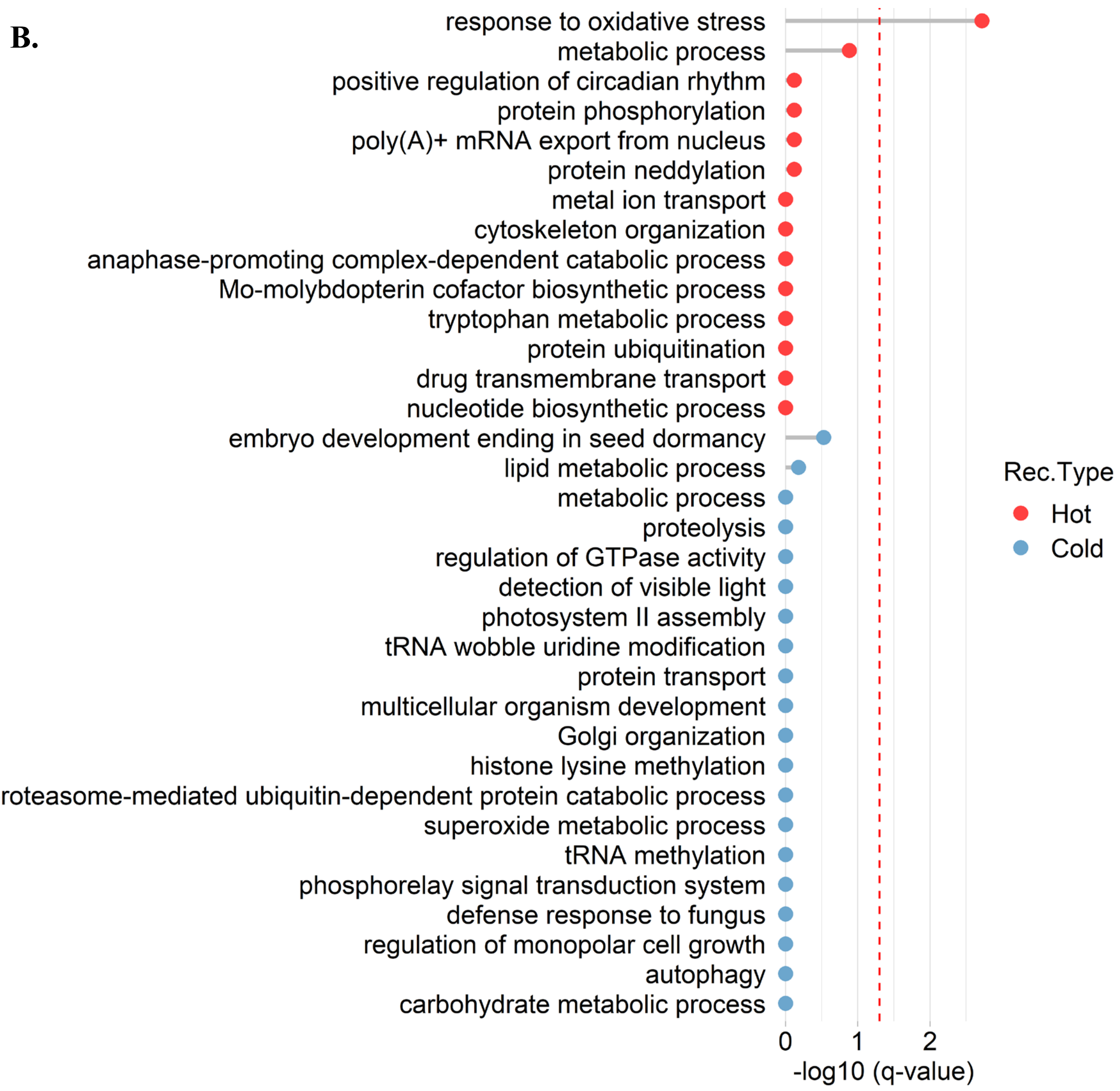

Figure 3.11 (continued)

Orthologous Copy Number Variation and Synteny

The numbers of orthologous genes within recombination cold spots and hot spots between *G. hirsutum* and *G. barbadense*, *G. hirsutum* and *G. tomentosum*, and *G. hirsutum* and *G. mustelinum* were compared against the genome wide numbers of orthologs between each respective species in order to assess if orthologous copy number variation (CNV) was significantly associated with low or high levels of recombination. Although there is variation in the number of orthologous copies within recombination types relative to their respective genomes, none of the comparisons were statistically significant ($p < 0.05$) (**Figure 3.12**). These results are surprising because non-allelic homologous recombination and non-homologous end joining are mechanisms attributed to the generation of CNVs and are associated with areas that have elevated recombination rates in both plants and animals [110-112]. Ortholog copy number was further categorized into three copy number types (graphically referred to as ortholog depth): one-to-zero (one gene present in non *G. hirsutum* species and ortholog not present in *G. hirsutum*), one-to-one (one ortholog present in non *G. hirsutum* species and one ortholog present in *G. hirsutum*), and one-to-many (one ortholog present in non *G. hirsutum* species and 2 or more orthologs present in *G. hirsutum*). Similar to the orthologous copy number variation analysis, there were no significant differences (p-value > 0.05) between the three orthologous depth types (**Figure 3.13D-F**) when compared to their respective genome wide distribution following a Chi-square goodness of fit test (**Figure 3.13A-C**).

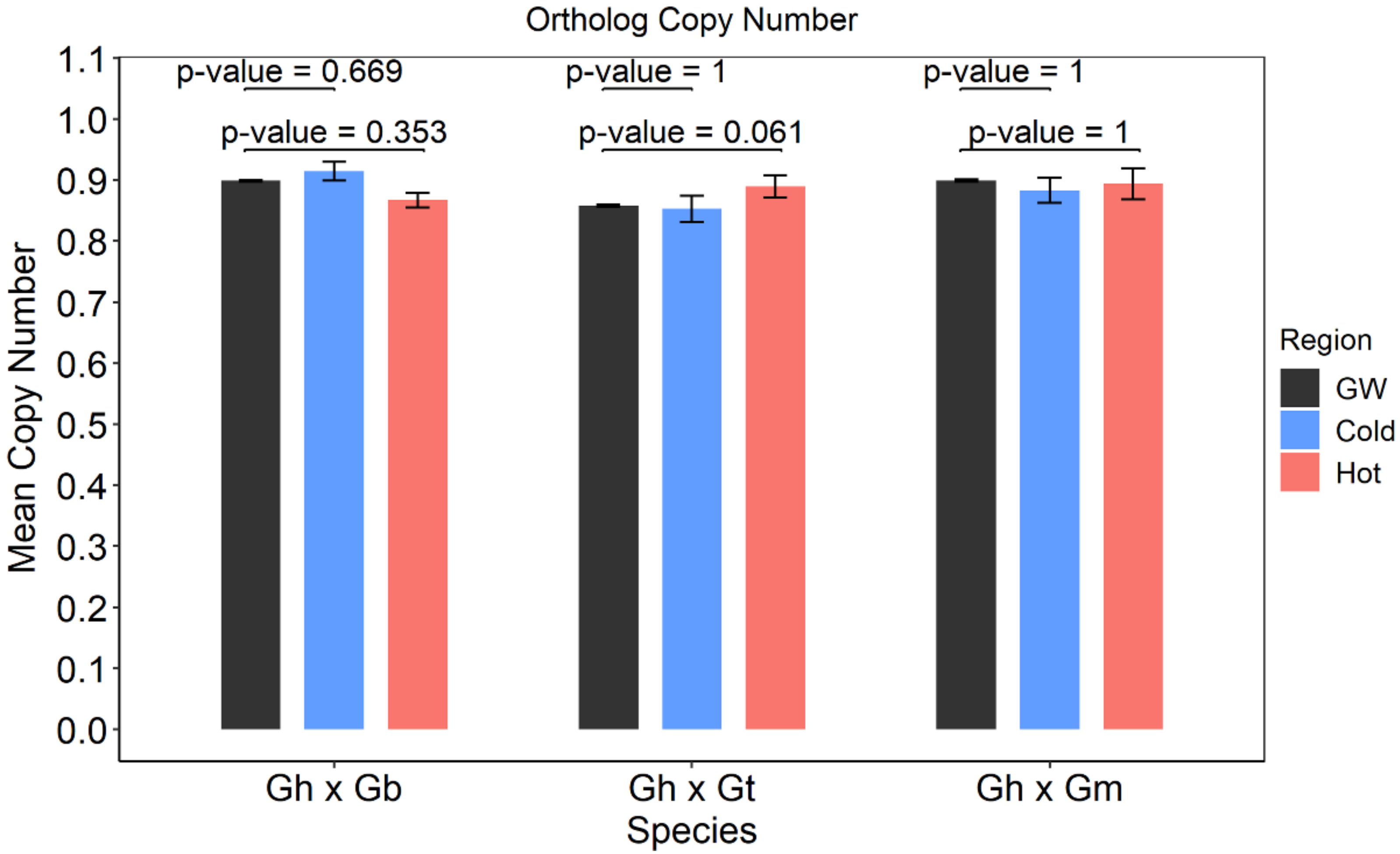

Figure 3.12 Ortholog copy number between recombination types. Average number of orthologs within recombination hot spots are shown in red while those from recombination cold spots are shown in blue. The average genome wide ortholog copy number is shown in black. Statistical test was conducted using a Wilcoxon rank-sum/Mann-Whitney test.

The primary transcript coding sequences for each orthologous pair were aligned to one another in order to assess synteny patterns between recombination cold spots and hot spots. There is variation in the synteny depth of recombination cold spots and hotspots compared to the genome wide synteny depth (**Figure 3.13D-F**), which is attributed to an increase in the number of one-to-many syntenic homeologous orthologs which can be seen in their respective karyotype maps (**Figure 3.14**). These results indicate that although there is neither a significant difference in orthologous CNV between recombination cold spots nor a difference in their proportion of copy number type, there is an increase in ortholog synteny involving their respective homeologs between *G. hirsutum* and *G. barbadense*, and *G. hirsutum* and *G. tomentosum* that are recombining at a high rate.

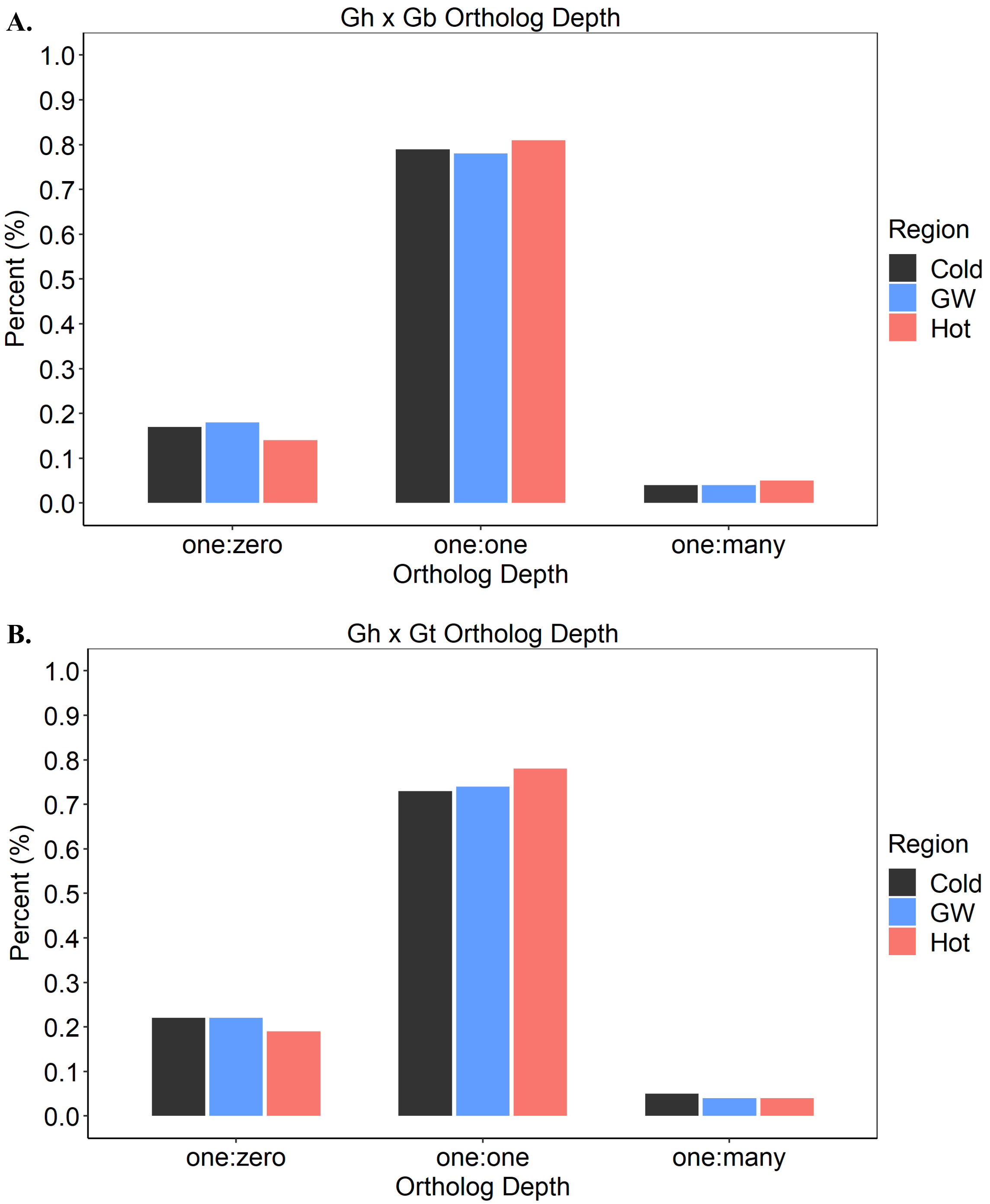

Figure 3.13 Proportion of copy number and syntenic depth between recombination types. Copy number between *G. hirsutum* and **A)** *G. barbadense*, **B)** *G. tomentosum* and **C)** *G. mustelinum* as well as syntenic depth between *G. hirsutum* and **D)** *G. barbadense*, **E)** *G. tomentosum* and **F)** *G. mustelinum* are shown. Orthologs within recombination hotspots are shown in red while those in recombination cold spots are shown in blue. The genome wide ortholog copy number and syntenic depth are shown in black.

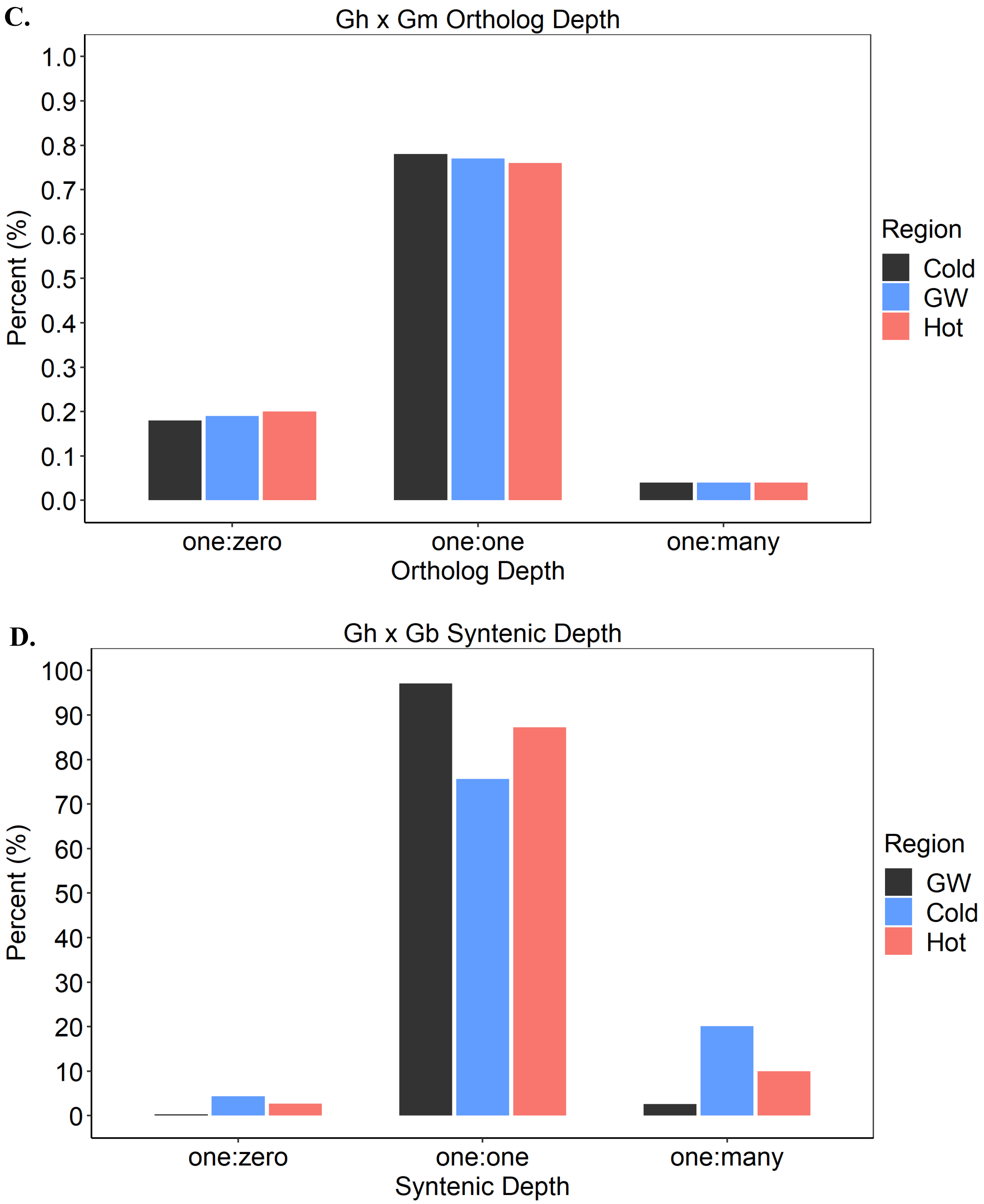

Figure 3.13 (continued)

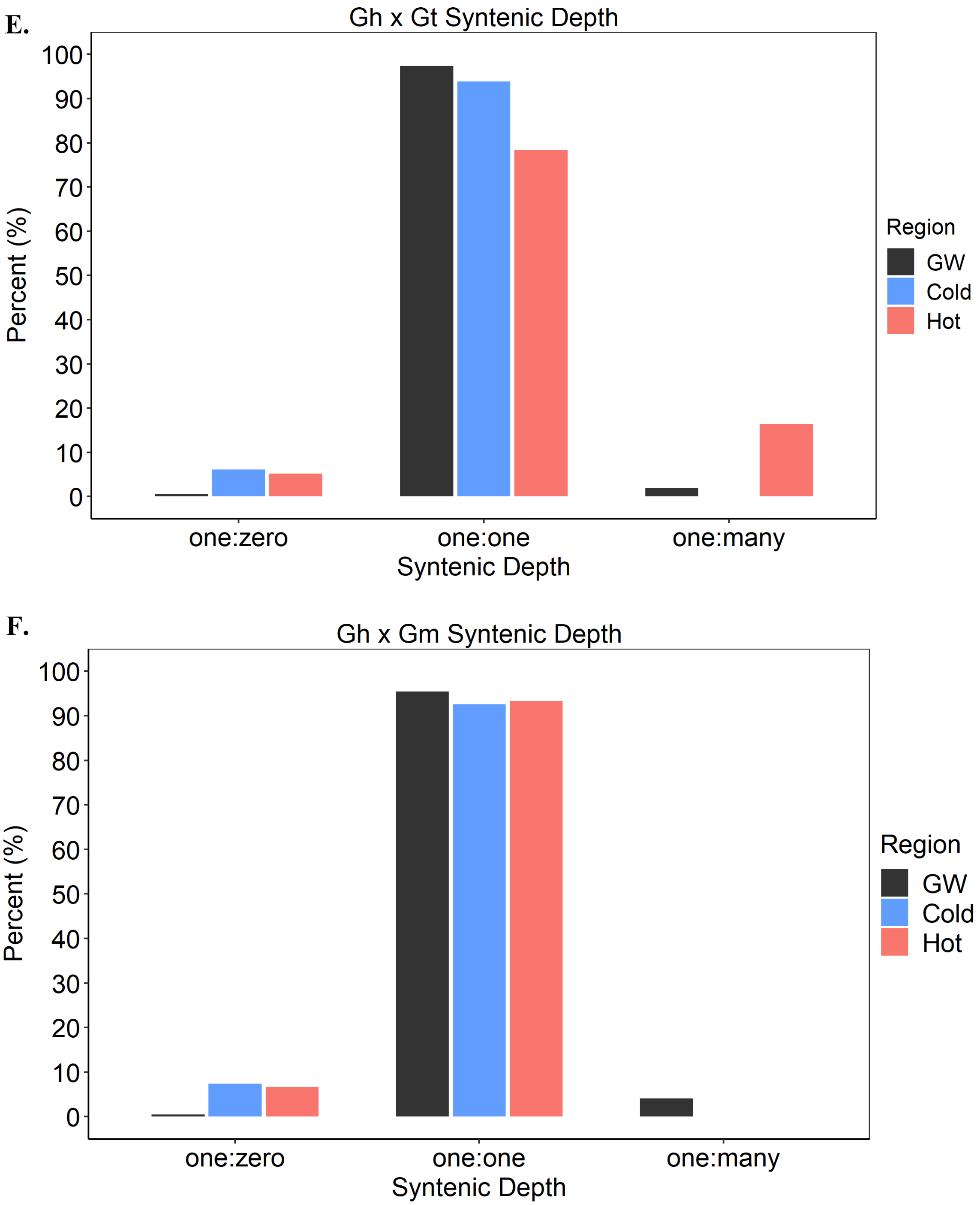

Figure 3.13 (continued)

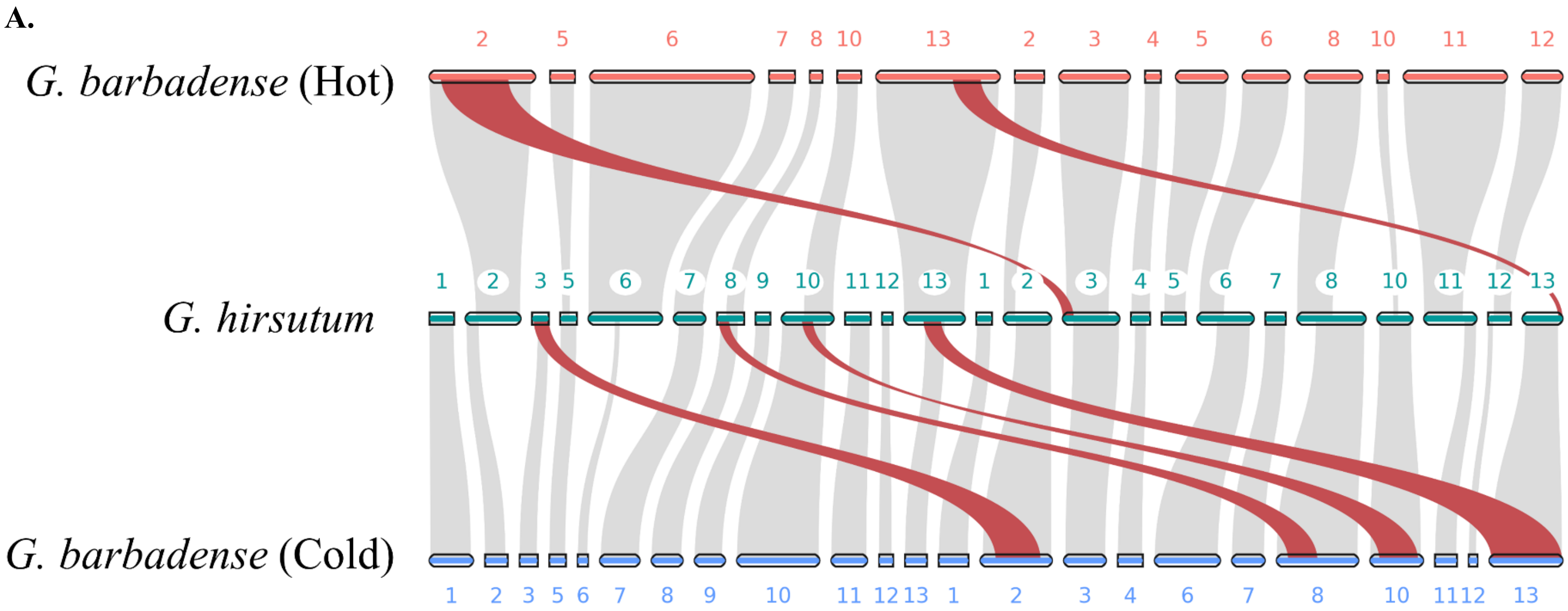

Figure 3.14 Synteny maps of recombination cold spots and hot spots. Synteny maps between *G. hirsutum* and **A)** *G. barbadense*, **B)**, *G. tomentosum*, and C) *G. mustelinum* are shown. Homologous synteny is represented through grey lines while homeologous synteny is represented through red lines.

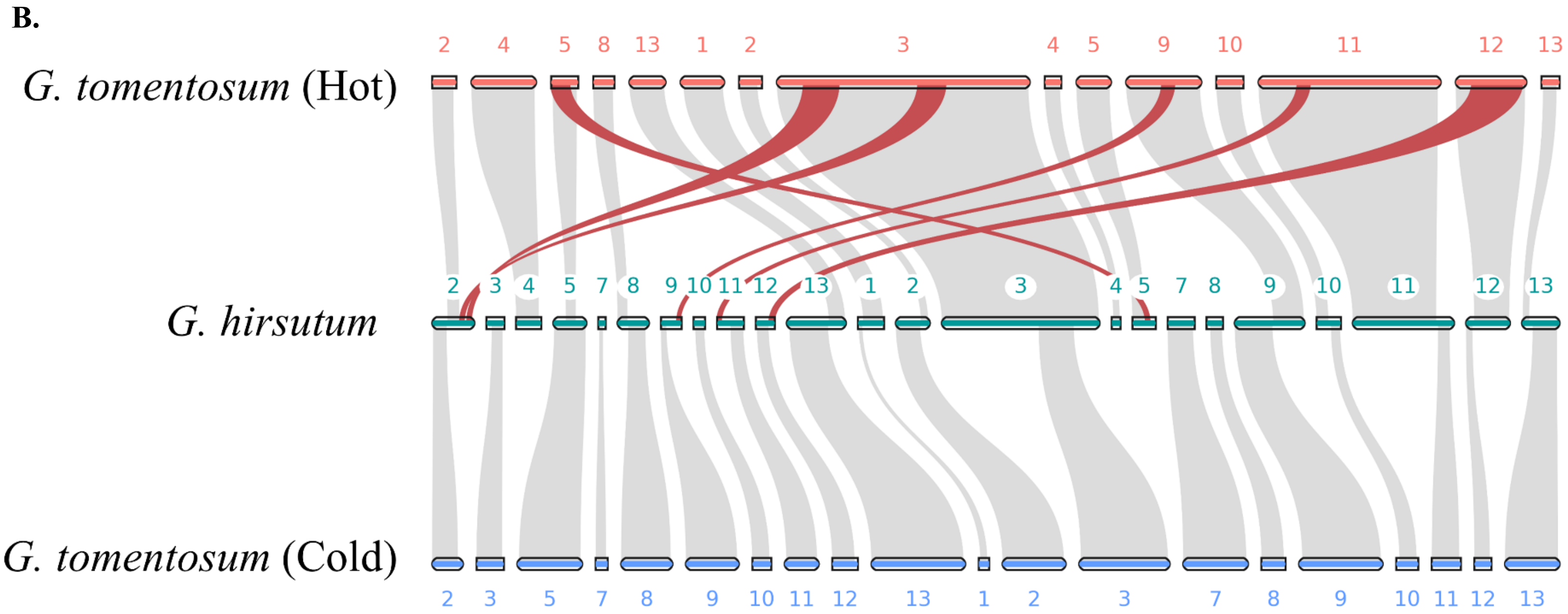

Figure 3.14 (continued)

C.

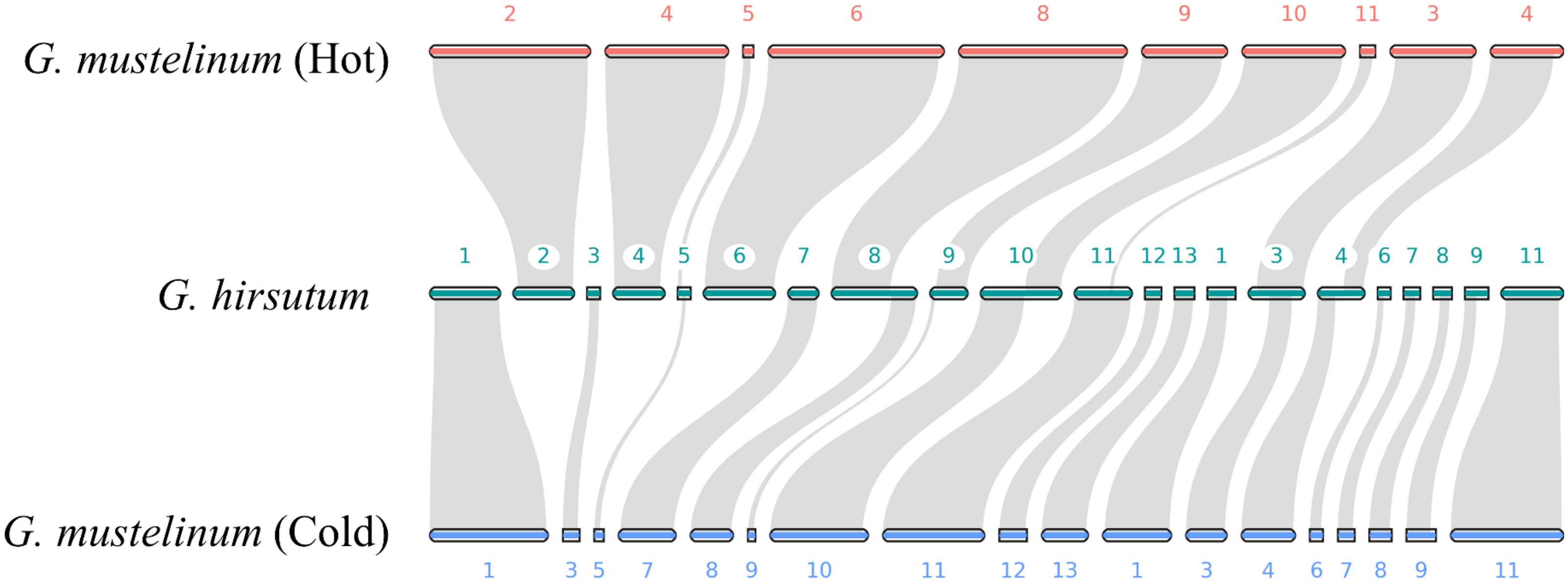

Figure 3.14 (continued)

Transposable Element Densities

Analysis of TE density within recombination cold spots and hot spots reveals that there is a significant negative correlation between genome-wide TE density and recombination hot spots in all of four of the mapping populations (**Figure 3.15A-D**). These results are not surprising as a negative correlation between TE density and recombination is seen in both plants and animals [47, 113, 114], but this relationship in *G. hirsutum*, or in the other three allotetraploid species, has not been demonstrated to date. There is a significant decrease in TE density between recombination cold spots and the genome wide TE density within *G. hirsutum* in two of the four mapping populations, but not within *G. tomentosum* and *G. mustelinum* (**Figure 3.15C-D**). This reveals an asymmetry in TE density between *G. hirsutum* and *G. tomentosum* and between *G. hirsutum* and *G. mustelinum* within these regions relative to their genome-wide TE densities. Interestingly, there are also select regions (outliers past the upper whisker) within recombination hot spots in all three of the interspecific populations that have a TE density that is greater than their respective genome-wide TE density revealing that these TE heavy regions are still recombining at elevated rates relative to their respective genomes. A previously published study [44] has revealed that there is a strong negative correlation between recombination rates and DNA methylation rates in *G. hirsutum*, *G. barbadense*, *G. tomentosum*, and *G. mustelinum*. There is also evidence that DNA methylation is a mechanism for the suppression and regulation of transposable elements in both plants and animals [90, 115, 116] and is capable of spreading past TE insertion sites leading to epigenetic regulation of adjacent chromatin [117-119]. The methylation of chromatin has also been shown to suppress crossovers at recombination hot spots through RNA-directed methylation [120]. The methylation states of highly recombinant and highly TE dense chromatin may be responsible for these outliers, e.g. heterozygous methylation states vs homozygous states, and should be investigated further given the role methylation serves in chromatin regulation and its impact on meiotic recombination.

A.

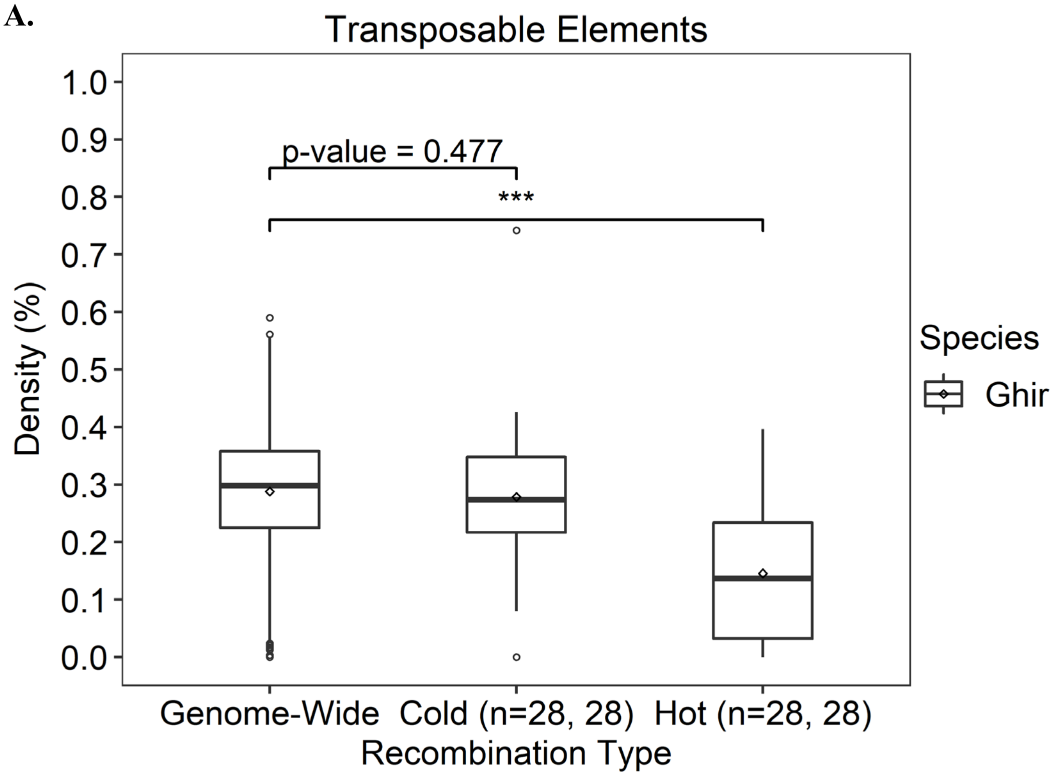

Figure 3.15 Transposable element densities within recombination types of four mapping populations. The **A)** consensus mapping, **B)**, GhxGb F_2, **C)**, GhxGt BC_1F_1, and D) GhxGm BC_1F_1 populations are shown. Statistical test was conducted using a Wilcoxon rank-sum/Mann-Whitney test with single (*), double (**), and triple (***) asterisk indicating a statistical significance levels of p-value < 0.05, <0.01, and <0.001, respectively.

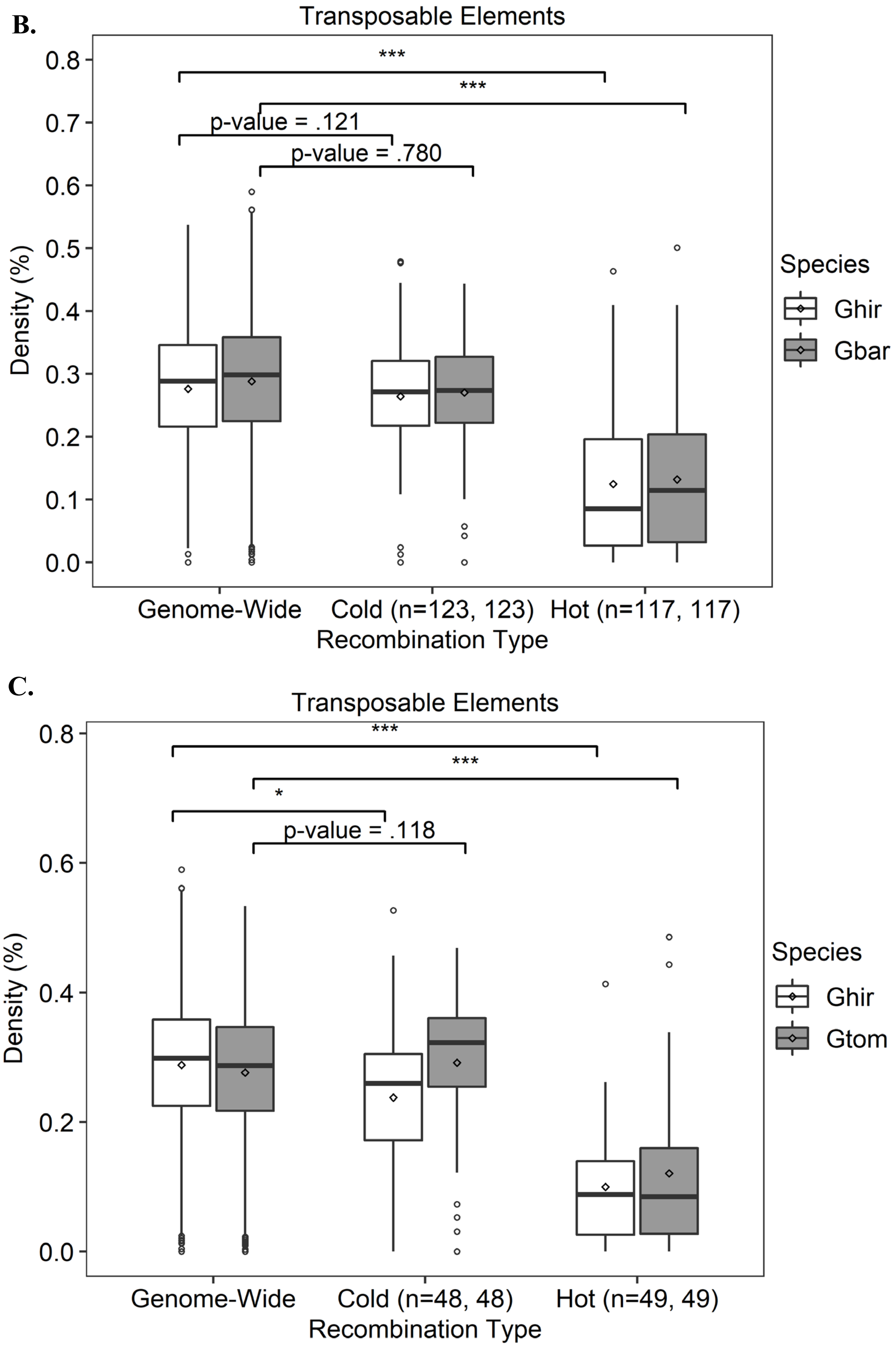

Figure 3.15 (continued)

D.

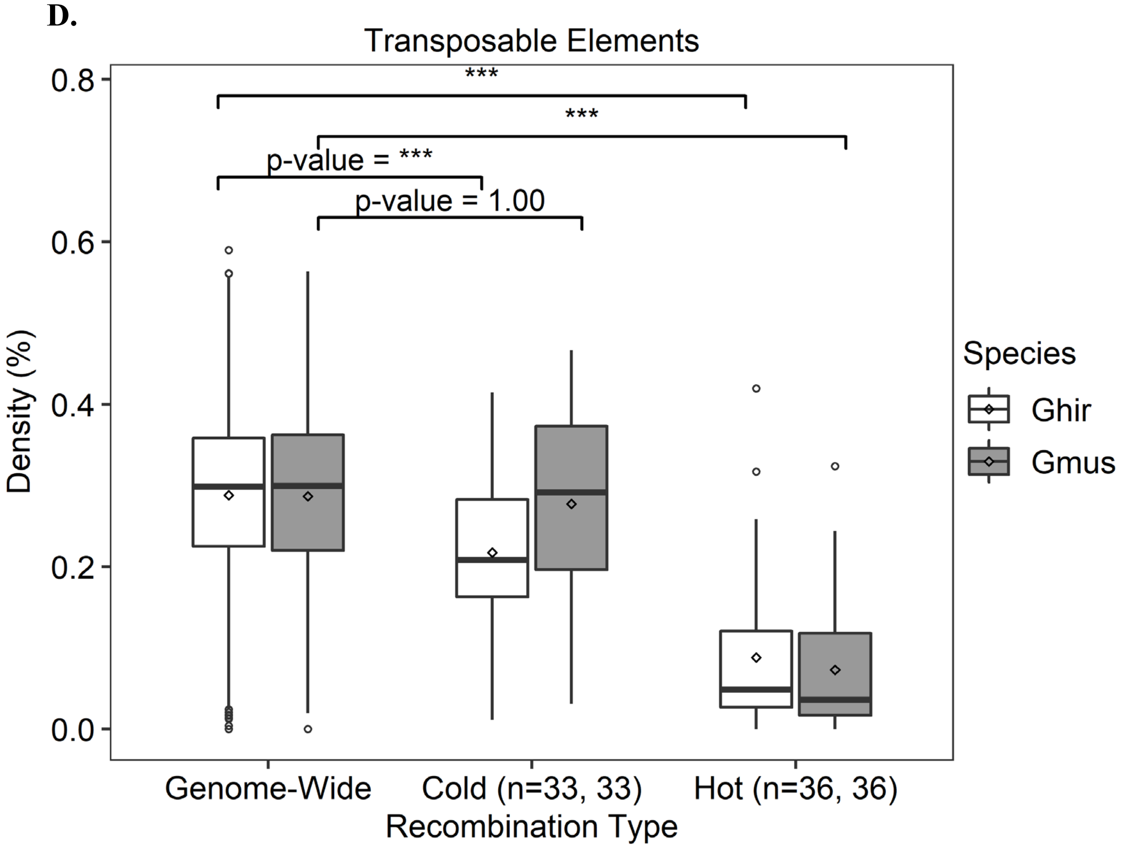

Figure 3.15 (continued)

Discussion

The recombination maps of three interspecific maps and one intraspecific map generated using the CottonSNP63K Array and mapped to long-read sequence assemblies have provided localized recombination rates which has allowed characterization of cis-features associated with meiotic recombination in four of the major allotetraploid cotton species. These recombination maps are also the first to be generated for interspecific populations consisting of cultivated Upland cotton and a wild cotton species. The distribution of recombination rates has also allowed for the analysis of recombination-specific chromatin that are homologous between species which was further facilitated by the use of the CottonSNP63K Array by allowing the direct comparison of homologous oligos containing polymorphic SNPs between the different cotton species. Comparisons of homologous chromatin within recombination hot spots and cold spots revealed strong conservation of chromatin lengths between different species implying accurate SNP mapping across the various species.

Gene enrichment (GE) analysis of recombination cold spots and hot spots allowed for the categorization of biological functions associated with these regions not just within cultivated Upland or Pima cotton, but within crosses involving undomesticated, wild cotton. The results reported in this study support the conclusions of Shen et al (2017) which published the first gene enrichment map of recombination cold spots and hot spots in Upland cotton and extends their analysis to three major allotetraploid cotton species. The functional categorization of genes reveals that genes within recombination cold spots are correlated with biological pathways associated with cell cycle stability while recombination hot spots are correlated with biological pathways associated with environmental stimuli. It is important to note that although a portion of the GE results in this study were statistically insignificant following multiple test correction, the algorithm by which the software topGO conducts GE takes into account the false-positive rate by traversing the GO hierarchy from bottom to top and deemphasizing GO terms higher in the hierarchy while simultaneously comparing the significance scores of parent and child nodes. This results in a reduction in the number of false positives at the cost of increasing false negatives [103, 104] so following the analysis with multiple test correction results in very conservative q-values and increases false negatives. As such, statistically significant GE results (p-value < 0.05) were reported along with their adjusted values (q-value < 0.05) similarly to Shen et al (2017) which reported unadjusted p-values in their results as well.

This study also allowed for the simultaneous GE analysis of homologous regions between *Gossypium* species, which to date, has never been done (**Figures 8-11**). Even though there was no statistically significant difference found in the number of orthologs between species within recombination cold spots and hot spots relative to their respective genomes (**Figure 3.12**), there was enough variation in the number or types of genes within these regions that resulted in different biological pathways being associated between these homologous regions. This suggests that there is sufficient genetic variation between species at these recombination cold spots and hot spots to trigger different biological pathways. These results provide evidence that introgression of germplasm from *G. barbadense*, *G. tomentosum*, and *G. mustelinum* can provide genetic variation to *G. hirsutum.*

As previously mentioned, there was no significant difference in the number of orthologs within recombination cold spots and hotspots between species relative to their respective genomes (**Figure 3.12**). There is also no significant difference in the proportions of ortholog copy number type implying that ortholog copy number is largely conserved within recombination type (**Figure 3.13A-C**). This may be a result of high DNA sequence homology between species at these recombination regions that is not seen elsewhere in the genome as CNVs tend to arise due to unequal crossing over between non-homologous sequences with high sequence fidelity. Although increased synteny between homeologous orthologous is seen between *G. hirsutum* and *G. barbadense*, and *G. hirsutum* and *G. tomentosum*, no major conclusions are drawn from these results (**Figure 3.14A-B**).

A major driver of the evolution of polyploid plants are transposable elements (TEs) as they can provide a novel source of genes, regulatory sequences, and structural variation [121]. Insertions of TEs also appear to be a major contributor to genome expansion in plants. Up to 85% of the genome content in maize (B73) is composed of TEs [122] while 90% of the wheat genome (*Triticum turgidum*) is composed of TEs [123]. The *Gossypium* genome has also undergone a threefold increase in size due to the accumulation of TEs resulting in the A subgenomes of *Gossypium* being approximately twice the chromatin size of the D subgenome [124]. The insertions of TEs also effects the epigenetic landscape of neighboring chromatin through a combination of methylation, RNA silencing, and histone modifications in order to silence the transcription and transposition of TEs [91, 125]. A strong association has also been established between TE density and recombination suppression [122, 126, 127], and in Upland cotton, a negative correlation between recombination and methylation has been shown [44]. The results of these study reiterate the association of TE density with recombination suppression. Recombination hot spots within all four of the mapping populations exhibit a significant decrease in TE density relative to their respective genome wide TE density (**Figure 3.15**). The TE density between species within recombination cold spots is also asymmetric between *G. hirsutum* and *G. tomentosum* and *G. hirsutum* and *G. mustelinum* relative to their respective genome wide TE densities (**Figure 3.15C-D**) implying that the TE landscape is not conserved between species within regions recombining at low rates.

The insertions of TEs may have an effect on the recombination rates within these regions as their epigenetic regulation can result in recombination suppression as previously mentioned. There are regions within recombination hot spots that have elevated TE densities relative to their genome wide density as seen in all three of the interspecific mapping populations. Modification to the epigenetic landscape that is not conserved between species, or differences in the types of TEs present may be responsible for these outliers as seen in alu-rich repeats in primates [128] or CACTA repeats in wheat [126] which exhibit a positive correlation with recombination. Further investigation into the influence of TEs on the epigenetic landscape in cotton may provide insights into mechanisms responsible for recombination rate regulation and may allow for artificial disassociations of linkages through epigenetic modification.

CHAPTER IV

INTROGRESSION OF DONOR *G. TOMENTOSUM* GERMPLASM INTO *G. HIRSUTUM* THROUGH THE DEVELOPMENT OF AN ADVANCED BACKCROSS POPULATION

Introduction

Advanced backcross populations are often used in breeding and breeding-related research programs to recover the elite parent's background genotype and to reduce or eliminate linkage drag while maintaining an allele, or alleles, of interest within the recurrent parent. This is typically employed when the recurrent parent is an elite genotype and hybridization with the donor genotype introduces unfavorable alleles. Backcrossing maximizes the recovery of the elite genotype and minimizes the risk of losing favorable alleles, but foreign genetic material from the donor parent can still remain. Therefore, extensive dilution through repeated backcrossing is typically needed to extensively "dilute" donor germplasm before any products with near-elite performance can be recovered. When coupled with wide-cross introgression, advanced backcross populations can serve as powerful breeding tools for introducing genetic diversity while maintaining near-elite performance in cotton while simultaneously providing insight into the recombinational landscape of hybrid-derived populations when marker-assisted selection is employed.

The development of a *G. hirsutum* and *G. tomentosum* CSSL panel will facilitate the introgression of wild, unadapted germplasm into a cultivated background by mitigating issues associated with linkage drag. Although each CSSL can contain more than one alien segment, the total of such segments in any single CSSL is expected be only a small fraction of the donor genome. The ultimate goal will be for the panel of CSSLs to span the entire donor genome, though the initial set of CSSLs may provide only partial coverage. Even so, the initial panel is expected to allow for [*i*] sensitive detection of quantitative trait loci (QTLs) associated with traits of interest or practical importance, through the prospective reduction of stochastic phenotypic "noise" and genetic variation, [*ii*] mapping or localization of the donor effect, as well as [*iii*] the prospective identification of linkage blocks (haplotypes) which hinder marker-assisted selection. We expect to use this CSSL panel to identify QTLs of interest at a further time and to associate these QTLs with SNPs. Comparison of haplotypic blocks to recombination landscapes for intra- and inter-specific linkage mapping populations led us speculate that wide-hybridization could facilitate disruption of haplotypic blocks and that could enable improvement of cotton in previously unforeseen ways. The CSSL panel will also allow for the assessment of haplotype block transmission across generations and provide insight on the efficacy of haplotype disruption through wide-cross introgression.

Materials and Methods

Development of G. tomentosum *Backcross Population*

A *G. hirsutum* x *G. tomentosum* BC_1F_1 population was created in the summer of 2012 in College Station, Texas by reciprocally crossing *G. hirsutum* x *G. tomentosum* F_1

individuals that were grown in the summer of 2011 at the same location. Twenty-seven BC_1F_1 individuals were selected for additional backcrossing to *G. hirsutum* as the male parent to create a BC_2F_1 population and twenty-seven BC_2F_1s were then backcrossed to the same genetic background in the summer of 2013 to create the BC_3F_1 population. In the summer of 2014, thirty-nine individuals each representing a family were backcrossed to develop 139 BC_4F_1 families. The BC_4F_1 were further backcrossed to develop 354 BC_5F_1 families in the summer of 2016. The BC_5F_1 population was then selfed in the summer of 2017 to develop the first BC_5S_1 individuals. The BC_5S_1 population was selectively selfed using genotyping data from thirty-nine BC3F1 individuals each representing one family in the summer 2018 to generate a BC_5S_2 population which was thereafter selfed in the summer of 2019 to generate the BC_5S_3. A subset of the BC_5S_2 population consisting of nine-hundred and two individuals were re-planted in the summer of 2020 alongside five-hundred and twenty-one BC_5S_3 individuals. All populations developed 2016 and after were germinated in Jiffy peat pellets (Kristiansand, Norway) in greenhouses prior to mechanical transplantation to a field in College Station, Texas.

Genotyping of the G. tomentosum *BC_3F_1 Backcross Population*

Young leaf tissue was collected from a panel consisting of thirty-nine individual BC_3F_1 individuals, each representative of a single family, and genotyped using the CottonSNP63K Array in 2017 at Texas A&M University. The genotype data were generated using Illumina's® GenomeStudio (v2.0) and converted to "ABH" format using a parent panel consisting of *G. hirsutum* and *G. tomentosum* allelic data. Only markers that were polymorphic between parents and that had less than 10% missing data were kept. The resulting BC_3F_1 genotype data were used to selectively propagate and self the *G. hirsutum G. tomentosum* advanced backcross population starting in the summer of 2018, as well as to identify theoretical introgression coverage of the *G. tomentosum* donor genome.

Haplotype and Introgression Analysis of Chromosome A08 of the BC_3F_1 Population

The processed BC_3F_1 genotype data were converted to PED/MAP format and used as an input for HaploView (v4.2). Haplotype block partitioning was similarly conducted using confidence intervals (CI). Allelic data of chromosome A08 in “ABH” format was compared to its haplotype structure to visualize individual lines for segregation within chromosome A08.

The BC_3F_1 was compared to the previously developed *G. hirsutum* x *G. tomentosum* BC_1F_1 linkage map to assess the proportion of introgression, determined by heterozygosity, in the BC_3F_1 relative to the BC_1F_1 mapping population.

Segregation Analysis of G. tomentosum *BC_1F_1 Mapping Population*

A high density *de novo* linkage map of a *G. hirsutum* x *G. tomentosum* BC_1F_1 mapping population of 85 individuals was developed as previously described in Chapter II. Testing for segregation distortion was conducted using JoinMap’s (v4.1) chi-square goodness of fit test across the entire linkage map. Markers that were significantly distorted (p-value < 0.01) based on their expected proportion were mapped to the JGI *G. hirsutum* v2.1 sequence assembly and converted to a BED formatted file using previously described methods. The BED file was intersected with the JGI v2.1 gene annotation file using bedtools intersect in order to determine gene identity within the significantly distorted regions. Genes found to be significantly distorted where then used to conduct gene enrichment analysis with topGO using the previously described methods. This procedure was identical to the gene enrichment methods conducted using the JGI *G. tomentosum* v1.0 sequence assembly. Marker segregation distortion and identity analysis was visualized using the R package ASMap (v3.6) [67].

Assaying of BC_5S_1 Population for Marker-Assisted Selection and Introgression

Young leaf tissue was collected from nine-hundred and twenty-nine BC_5S_1 individuals in the summer of 2018. Tissue was processed for DNA isolation using 96-Well Synergy™ Plant DNA Extraction Kits (Lebanon, NJ) per manufacturer's instruction. All DNA samples were assed for quality and quantity using BioTek®'s Cytation 5 Cell Imaging Multi-Mode Reader (Winooski, VT) at Texas A&M University. All DNA samples were then diluted to 10 ng/µL using and used for downstream PCR analysis.

DNA samples were dehydrated for one hour at 65°C and previously designed allelic-specific, simplex PCR Allele Competitive Extension (PACE) primer sets were used to determine allele dosage at individual loci. Each primer set consisted of two allele-specific forward primers and one common reverse primer that were designed using the CottonSNP63K Array ID oligos. PCR master mix was created using 2μL of Nanopure® purified water (Waltham, MA), 2μL of PACE reagent, and 0.056 μL of loci-specific primers for a total of 4.056 μL per 384 PCR plate well. PCR master mix was then added to the dehydrated DNA samples using a 384-well PCR plate. PCR samples were amplified using an ABI Veriti 384-well thermal cycler (Waltham, MA) for 44 cycles. Samples were then scanned using a PHERAstar® Microplate Reader (Ortenberg, Germany) and the output imaging data were analyzed using KlusterKaller software (Novata, CA) using default parameters. Control DNA samples including *G. hirsutum*, *G. tomentosum*, their respective F_1, and a no-template control (NTC) master mix sample without DNA were also included within the PCR plate using similar methods. Allelic composition of each loci was ultimately determined by comparing the fluorescent intensity of each allele to the respective controls.

Individual introgression sites that were heterozygous for *G. hirsutum* x *G. tomentosum* or homozygous for *G. tomentosum* were compared to the haplotype structure of the improved *G. hirsutum* cotton panel. Positive introgression sites were converted to a BED formatted file and intersected against the haplotype data of the improved cotton panel that was similarly converted to a BED formatted file.

Results

Linkage Map Construction

A high-density interspecific linkage map generated from 85 BC_1F_1 individuals between *G. hirsutum* and *G. tomentosum*. A total of 14,622 SNPs were mapped across

26 individual linkage groups resulting in a total genetic length of 4,045.82 cM (**Figure 4.1** and **Table 4.1**). There were a total of 1,157 uniquely mapped SNPs (bins) with an average number of 89 bins per chromosome within the A subgenome and a total of 1,112 bins with an average of 86 bins per chromosome within the D subgenome. The A subgenome had an average number of 537 SNPs per chromosome with an average chromosome length of 159.87 cM while the D subgenome had an average number of 588 SNPs with an average chromosome length of 151.35 cM. The increased genetic lengths within the A subgenome indicates increased recombination rates relative to the D subgenome.

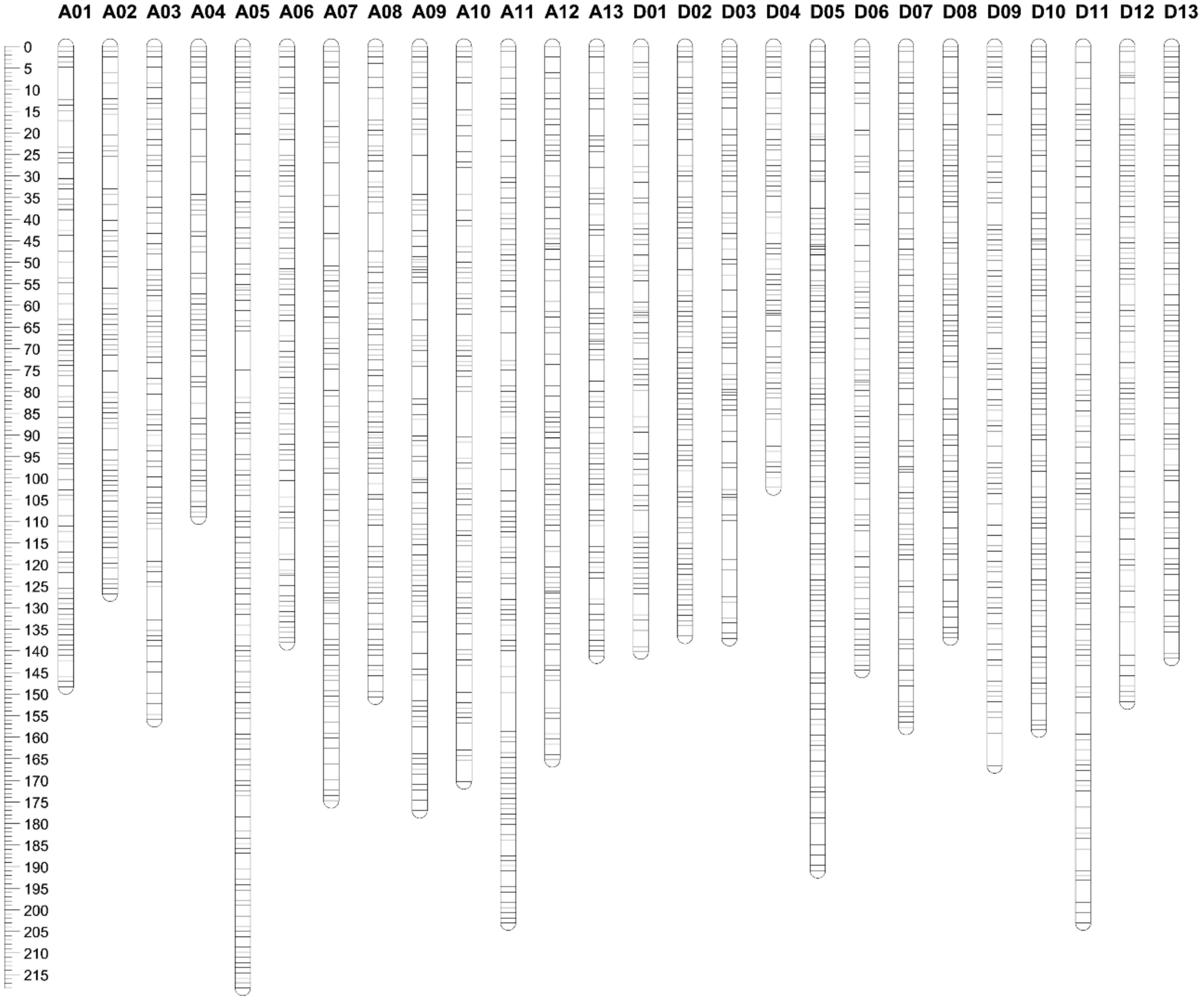

Figure 4.1 Interspecific linkage map of 26 cotton chromosomes. Genetic map was constructed using 85 BC_1F_1 generated from a *G. hirsutum* (TM-1) by *G. tomentosum*. The genetic distance (cM) is provided as a ruler left of the map.

Table 4.1 Summary of the high-density SNP map based on the BC_1F_1 population.

Chromosome	Length (cM)	SNP #	# of Unique SNPs	Average Interval (cM)	Largest Gap (cM)
A01	148.26	527	88	0.28	7.61
A02	126.84	374	72	0.34	7.61
A03	155.86	474	83	0.33	7.61
A04	108.95	262	58	0.42	7.61
A05	218.14	696	126	0.31	8.9
A06	138.06	402	88	0.34	6.25
A07	174.70	467	87	0.37	8.99
A08	150.67	894	93	0.17	8.99
A09	177.01	456	93	0.39	8.99
A10	170.32	579	83	0.29	10.42
A11	203.05	582	110	0.35	8.99
A12	165.20	573	94	0.28	6.25
A13	141.16	689	82	0.21	6.25
D01	140.17	626	77	0.22	6.25
D02	136.59	652	94	0.21	4.94
D03	137.12	391	17	0.35	8.99
D04	102.17	358	64	0.29	6.25
D05	190.88	1011	126	0.19	6.25
D06	144.55	661	85	0.22	6.25
D07	157.70	622	94	0.25	4.94
D08	136.96	761	83	0.18	4.94
D09	166.63	474	92	0.35	7.61
D10	158.34	551	100	0.29	3.66
D11	203.00	465	104	0.44	6.25
D12	151.82	563	88	0.27	7.61
D13	141.69	512	88	0.28	4.94
A_t	2,078.21	6,975	1,157	0.31	10.42
D_t	1,967.61	7,647	1,112	0.27	6.07
AD_t	4,045.82 cM	14,622	2,269	0.29	10.42

*A_t: A sub-genome

*D_t: D sub-genome

*AD_t: Both subgenomes

Segregation Distortion and Gene Enrichment Analysis

The chi-square goodness-of-fit test revealed significant deviations (p-value < 0.05) from the expected one-to-one allelic proportions within the BC_1F_1 mapping population (**Figure 4.2**). There were 1,481 total SNPs that had a p-value of less than five percent and 268 total SNPs that had a p-value of less than one percent. The distorted SNPs with a p-values of less than one percent were found across chromosomes A01, A03, A07, D02, D05, D07, D10, and D11.

The gene enrichment analysis of genes found within the significantly distorted (p-value < 0.01) regions revealed a total of 33 significantly associated (p-value < 0.05) biological pathways within *G. hirsutum* and 40 significantly associated (p-value < 0.05) biological pathways in *G. tomentosum* (**Figure 4.3)**. Twenty-three of the biological pathways were shared between both *G. hirsutum* and *G. tomentosum* with 10 pathways being unique to *G. hirsutum* and 17 pathways being unique to *G. tomentosum*. The increased number of biological pathways within *G. tomentosum* may reflect an increase variation in the number of genes or types of genes found within regions that are homologous to *G. hirsutum*.

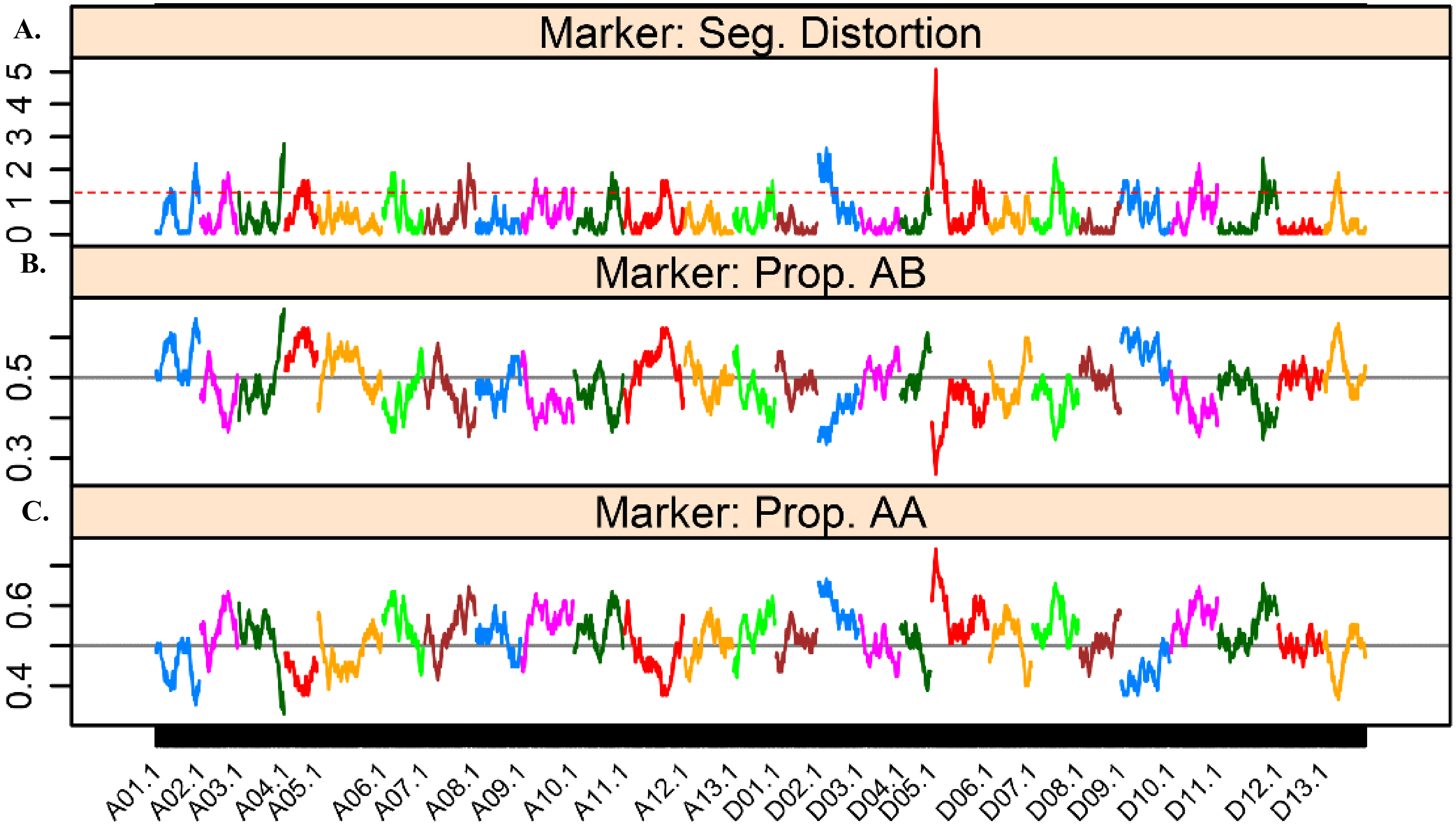

Figure 4.2 Segregation within the *G. hirsutum* x *G. tomentosum* BC$_1$F$_1$ mapping population. A) Markers under significant segregation distortion (p-value < 0.01 indicated by the dashed red line) as determined by a chi-square goodness-of-fit are shown. **B)** Proportion of markers heterozygous for *G. hirsutum* and *G. tomentosum* and C) and proportion of markers homozygous for *G. hirsutum* are shown.

A.

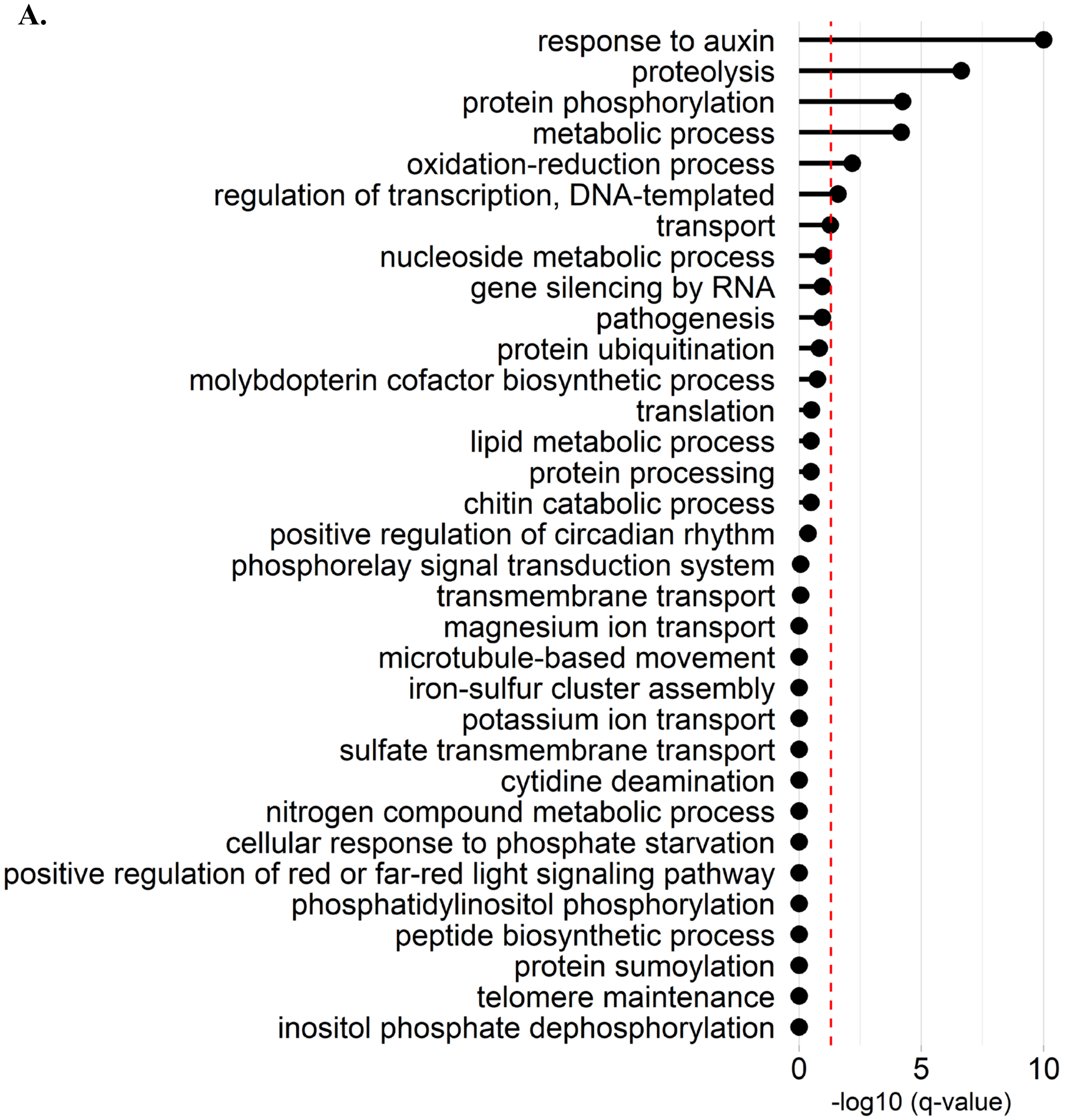

Figure 4.3 Functional categorization of genes under segregation distortion. A) Genes from the JGI *G. hirsutum* sequence assembly are shown in black while **B)** genes from homologous regions from the JGI *G. tomentosum* sequence assembly are shown in grey. All biological pathways shown have a p-value of less than 5% and biological pathways with a q-value (-log$_{10}$) of less than 5% are represented to the right of the red dashed line.

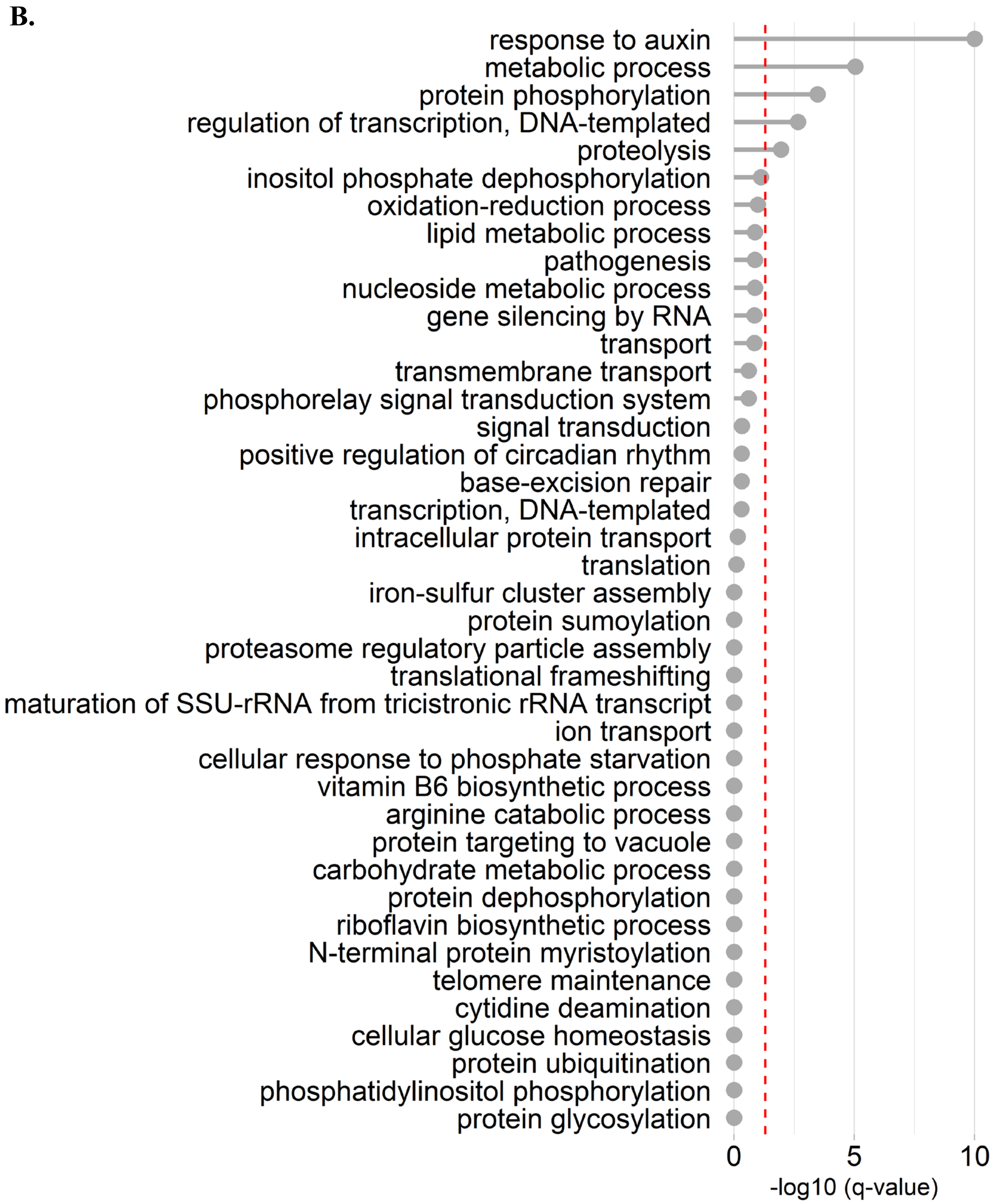

Figure 4.3 (continued)

Haplotype Structure of Chromosome A08 of the BC_3F_1 Population

A total of 39 *G. hirsutum* x *G. tomentosum* BC_3F_1 were genotyped using the CottonSNP63K Array and were retained following quality control (**Figure 4.4A**). A total of 15,930 SNPs were found to be segregating within the population which is an additional 1,308 SNPs compared to the BC_1F_1 mapping population. The haplotype structure of chromosome A08 was conducted using similar methods previously described resulting in the demarcation of 12 haplotype blocks (**Figure 4.4B**). A total of 1,003 SNPs were localized to chromosome A08, of which 658 SNPs were mapped within a large ~97.8 Mb haplotype block that is homologous to the large haplotype block within chromosome A08 of the improved cotton panel consisting of 257 improved *G. hirsutum* accessions. It is important to note that delineation of the haplotype structure of the BC_3F_1 population will be prone to inflated linkages due to the high linkage disequilibrium resulting from genotyping a very small population. A total of 6 recombination events (**Figure 4.4A**) are seen relative to the large haplotype block of the improved cotton panel indicating that introgression of genetic germplasm through wide-cross introgression has occurred and is not simply theoretical.

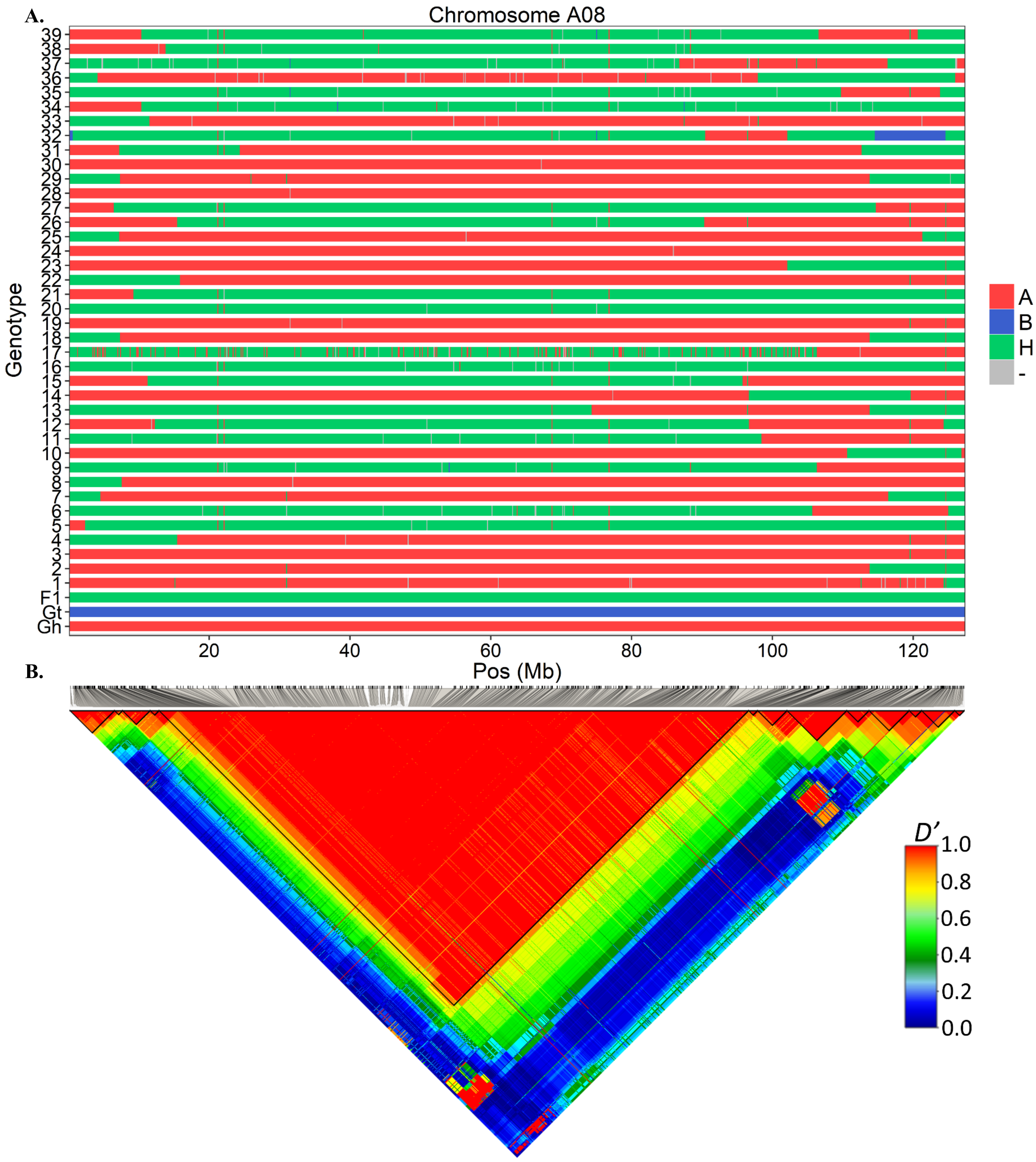

Figure 4.4 Allelic parentage and haplotype structure of chromosome A08 of a GhxGt BC$_3$F$_1$ population. A) The alleles homozygous for *G. hirsutum* (red), homozygous for *G. tomentosum* (blue), and heterozygous for *G. hirsutum* and *G. tomentosum* (green) are shown. Alleles with failed calls are shown in grey. **B)** Heatmap consists of equidistant tiles that indicate linkage disequilibrium as determined by a normalized coefficient of linkage disequilibrium (D') between pairs of markers. Black triangles within the heatmap demarcate haplotype boundaries Markers corresponding to SNP positions above the heatmap are congruent to the x-axis of the recombination map.

SNP Genotyping for Marker-Assisted Selection and Haplotype Disruption

A total of 929 DNA samples from a *G. hirsutum* x *G. tomentosum* BC_5S_1 chromosome segment substitution population were collected in the summer of 2018 and used for simplex SNP genotyping to detect introgressed chromatin from the *G. tomentosum* donor genome within the *G. hirsutum* background genome. Samples from the summer of 2018 were used as opposed to the more advanced 2019 BC_5S_2 due to a system infection of the BC_5S_2 population by a cotton leafroll dwarf virus which rendered a majority of the population sterile [129].

A total of 189 loci were SNP gentoyped using a PACE-PCR system across the A subgenome of the BC_5S_1 (**Figure 4.5** and **Table 4.2**), which constitued 49.6% of the total available SNP markers for introgression detection. A total of 69 (36.5%) of screened loci were positive for being homozygous for *G. tomentosum* or heterozygous bewteen *G. hirsustum* and *G. tomentosum* indicating succusfull detection of the donor genome within the background genome. Out of all of the primers that were tested, 68 (36%) failed to successfully amplify a loci or were incaple of distinguishing allele composition.

Loci that were succesfully identified for positive introgression for the donor *G. tomentosum* genome were mapped to the JGI *G. hirsutum* v2.1 sequence assembly using the CottonSNP63K Array IDs to assess the efficacy of haplotype block disruption. Postive introgression sites were converted to BED file format, with each BED coordinate consisting of a single base pair, and intersected to the haplotype structure of the improved cotton panel BED file, using the methods previously described, using the software bedtools intersect. A total of 32 haplotype blocks within the improved cotton panel returned a positive hit (**Figure 4.6**) of which 3 introgression intersected with haplotype blocks greater than 5 Mbs with an average haplotype block size was 2.15 Mbs. These results indicate that it is possible to disrupt large haplotype blocks within the improved cotton accesions using wide-cross germplasm introgression from wild *G. tomentosum*.

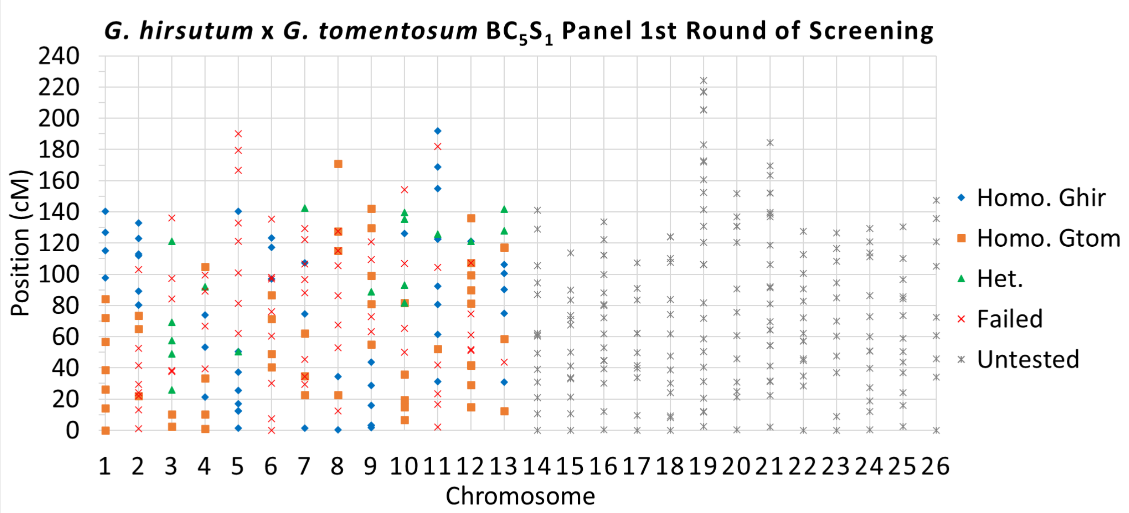

Figure 4.5 Screening for *G. tomentosum* chromatin introgression within a *G. hirsutum* genetic background. Screening was conducted using a PACE-PCR system for SNP detection on a GhxGb BC_1S_1 population. SNPs homozygous for *G. hirsutum* are shown as blue diamonds, SNPs homozygous for *G. tomentosum* are shown as orange squares, SNPs heterozygous for both species are shown as green triangles, and SNPs that failed to determine allele identity are shown as red crosses. Primers designed for SNP determination, but not tested, are shown as grey crosses. Position (cM) in shown on the y-axis while chromosome number (1-26) is shown on the x-axis.

Table 4.2 Summary of *G. tomentosum* introgression in TM-1 genetic background.

Allele	Count	Percent (%)
Homozygous for *G. hirsutum*	52	27.5
Homozygous for *G. tomentosum*	52	27.5
G. hirsutum x *G. tomentosum* F1	17	9.0
Failed	68	36.0
Total	189	

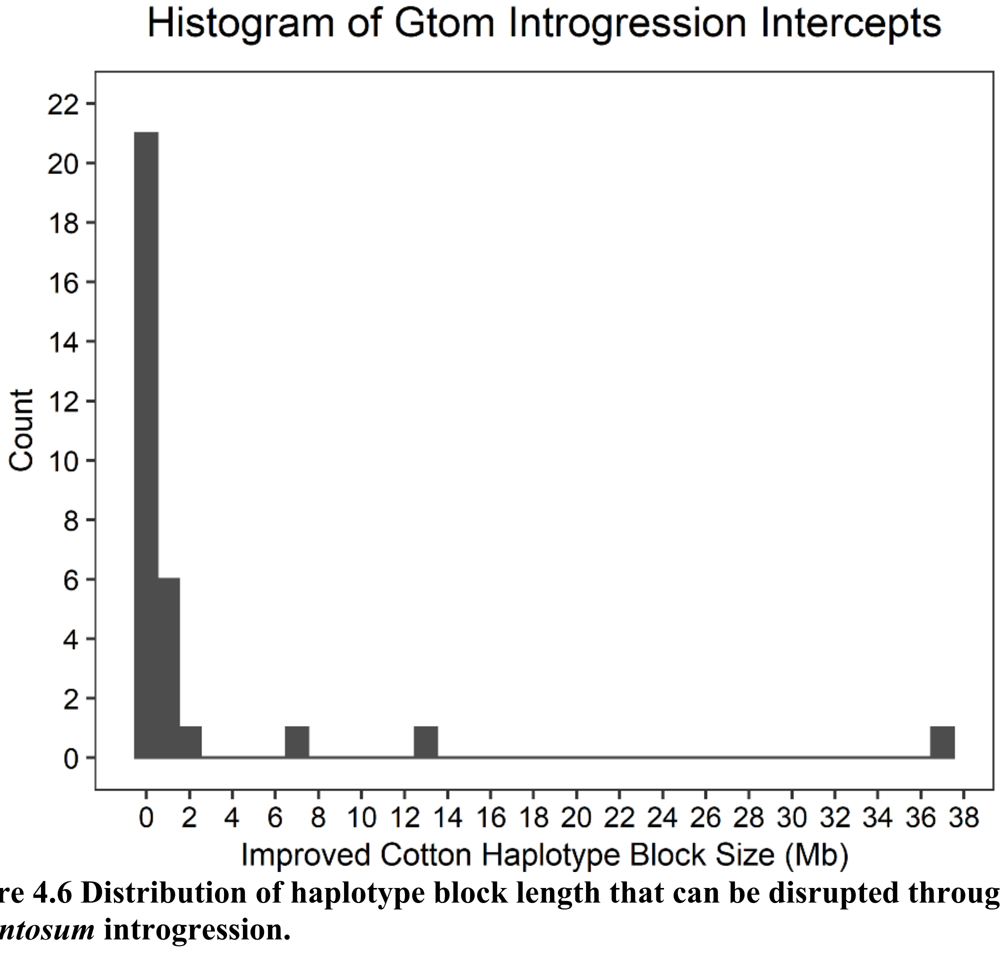

Figure 4.6 Distribution of haplotype block length that can be disrupted through *G. tomentosum* introgression.

Discussion

A *de novo*, high density *G. hirsutum* x *G. tomentosum* BC_1F_1 linkage map was developed using the CottonSNP63K genotyping array (**Figure 4.1**) and is a departure from previous linkage maps that utilized simple sequence repeats (SSRs), or microsatellites, for map construction [130, 131]. This linkage map is also the highest saturated map to date and surpasses the total number of SNPs detected from recently developed *G. hirsutum* x *G. tomentosum* BC_2F_2 mapping population that utilized a genotyping-by-sequencing approach [132]. The improvement is important for the interspecific introgression of wild germplasm from the Hawaiian cotton species *G. tomentosum* into Upland cotton to improve the latter's phenotypic performance through selective genetic expansion. The use of backcross introgression to disperse the *G. tomentosum* donor genome into a TM-1 background genome will also help mitigate the deleterious effects of crossing a cultivated crop with a wild species.

One of these negative consequences of wide-cross introgression is the introduction of poorly performing agronomic traits into the improved genome through linkage drag. Development of a chromosome segment substitution population (**Figure 4.5**) will help mitigate that effect by separating the donor genome across multiple *G. hirsutum* lines allowing for greater specificity in selecting *G. tomentosum* genes within the *G. hirsutum* genome relative to other wide-cross population such as those developed into F_2 hybrids or recombinant inbred lines (RILs), which generally result in 1:2:1 and 1:1 genome content ratios, respectively.

The potential disruption of markers under high linkage disequilibrium via wide-cross germplasm introgression has also been theoretically demonstrated within the BC_1F_1 mapping population, as demonstrated in an earlier chapter (**Figure 2.11**), and practically within the BC_5S_1 CSSL population. Assaying the BC_5S_1 CSSL population resulted in the detection of 31 introgression sites from *G. tomentosum* that correlate to haplotype blocks within the improved cotton panel (**Figure 2.5**), referenced in an earlier chapter, of which 5 haplotype blocks are greater than 5 Mb in size (**Figure 4.6**). Although selected introgression of wild germplasm does not guarantee improvement of agronomic traits within Upland cotton, it does allow for the potential exploration of novel phenotypic traits which may be beneficial in one capacity or another.

The non-mendelian transmission of chromatin from the F_1 to the BC_1F_1 may also provide insight into allelic interactions and potential retention of the *G. tomentosum* donor chromatin within Upland cotton. Overall, 52.2% of the genotyped alleles in the population were homozygous for *G. hirsutum* while 47.8% were heterozygous for both genomes **(Figure 4.2)** indicating a strong adherence to overall expected segregation ratios at this generation. The proportion of homozygous alleles for *G. hirsutum* that were significantly distorted with a p-values of less than 5% was 55.2% compared to significantly distorted heterozygous alleles at 44.8%. The biggest difference in proportion between both allele types was seen in markers that were significantly distorted with p-values of less than 1%. The proportion of distorted homozygous *G. hirsutum* alleles was 60.71% while the proportion of distorted heterozygous alleles was 39.3% at p-values of less than 1%. These results indicate that there is a bias against the heterozygous state at these specific loci. The gene enrichment analysis of biological pathways associated with genes within regions under significant segregation distortion revealed responses to auxin in both species, but no major processes that could be correlated segregation distortion were identified [133]. The overall retention of the donor genome and the effects of haplotype incompatibilities between the two species genomes may not be fully assessed until completion of the CSSL panel is achieved and is fully genotyped.

Overall, the development of a high-density SNP-based linkage map coupled with the creation of a chromosome segment substitution line will facilitate the concomitant introgression and dissection of a donor genome, and allow for replicated experiments needed to identify important quantitative variation for important traits of Upland cotton. Such quantitative effects are typically not visible at the donor species or plant level, due to overwhelming under-performance for domesticated traits and also strong epistatic effects.

CHAPTER V

DEVELOPMENT OF NOVEL GERMPLASM FOR INSTIGATING ECTOPIC RECOMBINATION WITHIN *G. HIRSUTUM* THROUGH GENETIC HEMIZYGOSITY

Introduction

Despite their common ancestry and overall sequence relatedness, the A and D subgenomes of *G. hirsutum* are considered to regularly relegate recombination to homologs and avoid reciprocal genetic recombination with each other, i.e., to avoid homeologous reciprocal recombination. In hexaploid wheat (*Triticum aestivum* L. 2n = 6x = 42), the avoidance or suppression of homeologous recombination and allowance of only homologous chromosomes to pair was found to be largely controlled by the Pairing Homeologous 1 gene (*Ph1*) [134]. Evidence of a comparably qualitative effect on cotton meiotic pairing and recombination has been lacking [135]. In *Gossypium* AD haploids and an AD hybrid between A2- and D2-genome diploid species, pairing between homeologs was evident at pachynema but diminished greatly by metaphase I, which led to the hypothesis that differences in repetitive sequence content and genome size could account for lack of homeologous recombination. However, advances in our understanding of the meiotic recombination process now recognize the homology search at leptonema, i.e., double-strand break (DSB) formation, as key a seminal event that leads into homologous chromosome pairing and the initiation of homologous recombination, including crossing over [136]. Evidence interpreted to indicate evolutionarily significant amounts of gene conversion between A and D gene homeologs has been reported for extant AD-genome *Gossypium* species [137]. Either or both of these observations could suggest that AD pairing can occur, but that "maturation" of any AD DSBs into crossovers is somehow precluded. In brewer's yeast (*Saccharomyces cerevisiae*), slightly diverged DNA sequences are sufficient to prevent recombination from occurring through the mismatch repair system (MMR) [136, 138]. It seems plausible that evolutionary divergence of 5-10 million years [3] between the A and D subgenomes in cotton could act as a strong barrier to reciprocal exchange leading to crossovers, possibly without precluding non-reciprocal exchange (conversion). Despite this, the asymmetric evolution of homeologous genes has been implicated in fiber deployment and environmental adaptation [139] providing evidence for the consequence of intergenomic exchange of genetic material between the subgenomes. Thus, experimental analysis is desirable given that homeologous recombination in AD cotton is still far from fully understood. It is reasonable to infer that understanding the effects of homeologous genes and CNVs in cotton can be facilitated by inducing homeologous recombination in cotton and examining the genetic and phenotypic consequences.

It is reasonable to infer that distributions and perhaps rates of homologous recombination can be altered and homeologous A-D intergenomic recombination in allotetraploid cotton can be induced through cytogenetic manipulations, specifically the use of hypoaneuploids to create chromosomes or segments that are hemizygous. The *G. hirsutum* inbred line, 'Texas Marker-1' (TM-1) was used as a recurrent parent to develop isogenic hypoaneuploid monosomic stocks ("H" – missing its homologous partner, e.g., H01 has just one, not two copies of chromosome-1). The monosomic stocks (M) will be crossed as females with euploid TM-1 isolines that are homozygous for specific reciprocal translocations (TT) involving the homeolog of the monosome (**Figure 5.1A**). Multiple reciprocal translocation homozygotes (**Table 5.1**) will be used per monosome to increase the probability and coverage of homeologous recombination. The resulting M/TT-cross progeny will be heterozygous for the translocation and should segregate for euploid (most) and maternally transmitted monosomy (some). Translocation heterozygotes of cotton fairly often generate maternally transmissible adjacent-1 and adjacent-2 meiotic products, including segmental duplication-deficiencies, as well as occasional numerical non-disjunction products, which can lead to primary monosomic and tertiary monosomic progeny. So, the selected monosomic translocation heterozygous (MTH) plants will be crossed as female to wild-type TM-1 (**Figure 5.1B**) to recover maternal products that feature co-recovery of the maternal monosomy and a new translocation-derived segmental duplication-deficiency or monosomy (primary or tertiary), where the translocation-caused deficiency involves the homeolog of the monosome **(Figure 5.1C**). The selected double-hemizygote (monosomic for one chromosome, and segmentally monosomic for part of its homeolog) single-translocation line will be screened and characterized for potential homeologous recombination and copy number variation (**Figure 5.1C-D**).

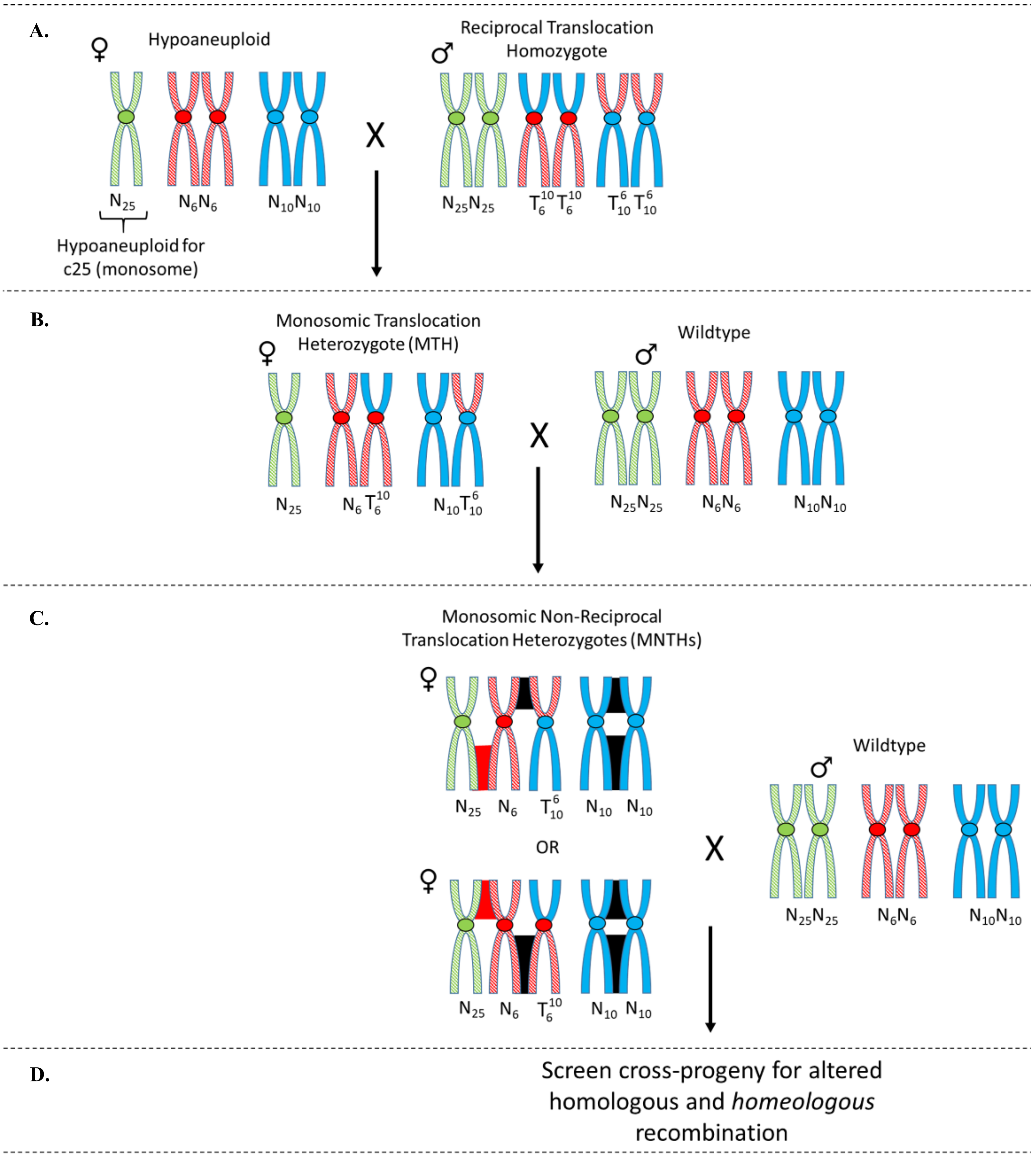

Figure 5.1 General breeding scheme to alter homologous and potentially induce homeologous recombination. **A)** An individual hypoaneuploid for chromosome 25 (monosomic for N_{25}) is crossed as a female to a euploid individual that is homozygous for reciprocal translocation **B)** The selected cross progeny is monosomic for chromosome 25 and heterozygous for the paternal reciprocal translocation. When crossed to a wild-type individual, the MTH's monosomy and translocation heterozygosity will lead to formation of *monosomic non-reciprocal translocation heterozygotes* (MNTHs). **C)** Zygotic MNTHs deficient for different segments of chromosome-10 and chromosome-6. Segments potentially amenable to homologous recombination are shown in black, whereas those potentially amenable to homeologous recombination are shown in red. Individuals are crossed with wildtype *Gossypium* spp. **D)** Progeny are to be screened for recombination events.

Materials and Methods

Growing Reciprocal Translocation Homozygotes and Hypoaneuploid Plants

Seed consisting of twenty-four lines, ten replicates each, derived from isogenic *G. hirsutum* (BC_5S_n) reciprocal translocation homozygotes that had been previously developed [140] were planted in Jiffy peat pellets (Kristiansand, Norway) in a greenhouse over the Fall of 2017 in College Station, Texas. Seedlings that successfully germinated were transplanted to larger pots and were allowed to grow to maturity. Three additional hypoaneuploid lines monosomic for either chromosome 10, chromosome 20, and chromosome 25 that were fully matured were procured from greenhouse plant stocks.

Seed consisting of ten lines, one-hundred replicates each, derived from ten separate isogenic *G. hirsutum* (BC_5S_n) hypoaneuploid individuals were grown in a greenhouse over the Fall of 2017 in College Station, Texas. Each line was descendant from an individual that was hypoaneuploid for chromosome 1 (H01), 2 (H02), 4 (H04), 6 (H06), 7 (H07), 10 (H10), 16 (H16), 17 (H17), 20 (H20) , or 25 (H25). Young leaf tissue was collected from seed that successfully germinated for DNA isolation using a Machery-Nagel Plant Nucleo-spin kit (Pennsylvania) according to the manufacturer's instructions. DNA samples were analyzed for quality and quantity using a Denovix DS-11 spectrophotometer (Wilmington, DE) and diluted to 10 ng/µL.

Copy Number Variation Analysis of Hypoaneuploids

Diluted DNA samples were dehydrated using the previously described method and screened for whole chromosome copy number variation, *monosomy* or *hypoaneuploidy*, using a previously developed comparative locus assay (CLA) consisting of two sets of primer pairs for a total of four primers. Primers for the CL assay were designed using oligos from the Cotton63KSNP Array and each pair consisted of a non-allelic specific forward primer that is conjugated with a fluorescent chromophore (VIC or HEX) and a common reverse primer. The secondary primer pair similarly consisted of a non-allelic specific forward primer with an alternate chromophore and a common reverse primer. One pair of primers is designed to amplify a locus that is on the putative monosomic chromosome while the second pair is designed to amplify a locus on a control, euploid chromosome.

PCR master mix was created using 2µL of Nanopure® purified water (Waltham, MA), 2µL of PACE reagent, and 0.056 µL of loci-specific primer specific for one chromophore, and 0.056 µL of loci-specific primer specific for a separate chromophore for a total of 4.112 µL per 384 PCR plate well. PCR master mix was then added to the dehydrated DNA samples using a 384-well PCR plate. PCR samples were amplified using an ABI Veriti 384-well thermal cycler (Waltham, MA) for 24 cycles. Samples were then scanned using a PHERAstar® Microplate Reader (Ortenberg, Germany) and the output imaging data were analyzed using KlusterKaller software (Novata, CA) using default parameters. Control samples consisting of *G. hirsutum* DNA and a NTC sample were similarly included. PCR samples were amplified again for one cycle and similarly scanned. This process was repeated until 44 complete PCR cycles were completed resulting in twenty amplification points per primer pair. Amplification data were normalized using the control *G. hirsutum* and NTC amplification data and the normalized, relative fluorescence units (RFUs) between chromophores was compared to determine copy number variance at the chromosome level. Putative hypoaneuploids were re-screened with two additional replicates for a total of three replicates to validate hypoaneuploidy.

Generating Monosomic Translocation Heterozygote (MTH) F_{1}s

Hypoaneuploid individuals that were successfully identified as being monosomic for their respective chromosome were used as females and crossed to a plant that contained a homozygous reciprocal translocation that was homeologous to the chromosome that was monosomic in the female (**Figure 5.7A** and **Table 5.1**). Crosses were conducted in a greenhouse during the spring of 2018, the summer of 2018, the fall of 2018, and in the Spring of 2019. The latter three crosses were conducted following a population loss due to inclement weather during the spring of 2018, fungal infection of F_1 bolls during the summer of 2018, and thrip destruction of F_1 plants during the winter of 2018. Crosses from the Spring of 2019 were successful in generating F_1 individuals for further hypoaneuploid screening (**Figure 5.7B**).

Screening for MTH F_1s Through Seed Extraction

The F_1 seeds resulting from the hypoaneuploid by reciprocal translocation homozygote crosses were screened for transmission of a monosomic chromosome from their respective hypoaneuploid parent using a seed extraction protocol [141]. Seeds were sanded on the opposite end of the funiculus using a Dremel sander (Dremel #407) and placed within modified 96-well PCR plates. The sanded portion of the seed was drilled with an engraving Dremel (Dremel #107) to approximately one millimeter to extract endosperm tissue. The modified 96-well PCR plates were placed within non-modified 96-well PCR plates and centrifuged for three minutes at four-thousand relative centrifugal force (RCF) to transfer the endosperm tissue into the regular PCR plate. The extracted seed tissue was then used for DNA isolation using a previously published sodium hydroxide protocol [141]. DNA samples were not diluted and were screened for hypoaneuploidy using the previously described CLA protocol.

Seeds that were putatively identified for hypoaneuploidy were planted in peat pellets in a greenhouse during the Fall of 2019. Young leaf tissue was collected and used for DNA extraction using a Synergy™ 2.0 Plant DNA Extraction Kit (Lebanon, NJ) per manufacturer's instructions. DNA samples were assessed for quantity and quality using a Denovix DS-11 spectrophotometer (Wilmington, DE) and diluted to 10 ng/µL. Diluted DNA samples were dehydrated and used for hypoaneuploid screening using the previously described CLA protocol. The CLA screening was repeated multiple times, with novel primer pair combinations, until at least three unique primer combinations with three replicates each resulted in positive hypoaneuploidy. Individuals that were positive for hypoaneuploidy were transferred to large growing pots and allowed to grow to maturity.

Results

Development and Recovery of Parental Lines

Plant material for instigating ectopic homeologus recombination through genetic hemizygosity is presented in **Table 5.1**. The total number of reciprocal translocation heterozygotes that successfully germinated was 120 (50%) out of the 240 plants that were initially planted. These 120 individuals represented 12 unique lines, or distinctive phenotypes, and where crossed to their respective hypoaneuploid partner as males. The reduced germination rate was likely due to the age of the seed as a majority of the seed was more than 5 years old at the time of planting.

A total of 1,000 seed descendent from hypoaneuploid lines were planted over the fall of 2017 (**Table 5.2**). A total of 774 seeds successfully germinated and were sampled for copy number variation. Screening of DNA from 291 individual plants resulted in successful amplification of targeted loci using the CLA-PACE-PCR system (**Figure 5.2**). Out of these, 5 plants resulted in positive hypoaneuploidy following additional validation screening with multiple replicates (**Figure 5.3**). A small subset of the total population was ultimately screened due to mechanical damage inflected on young leaf tissue by various pests within the greenhouse and slow germination of these lines. An additional 4 hypoaneuploid lines, two H10s, one H20, and one H25, were procured from greenhouse stocks and were also used as female parents for crossing.

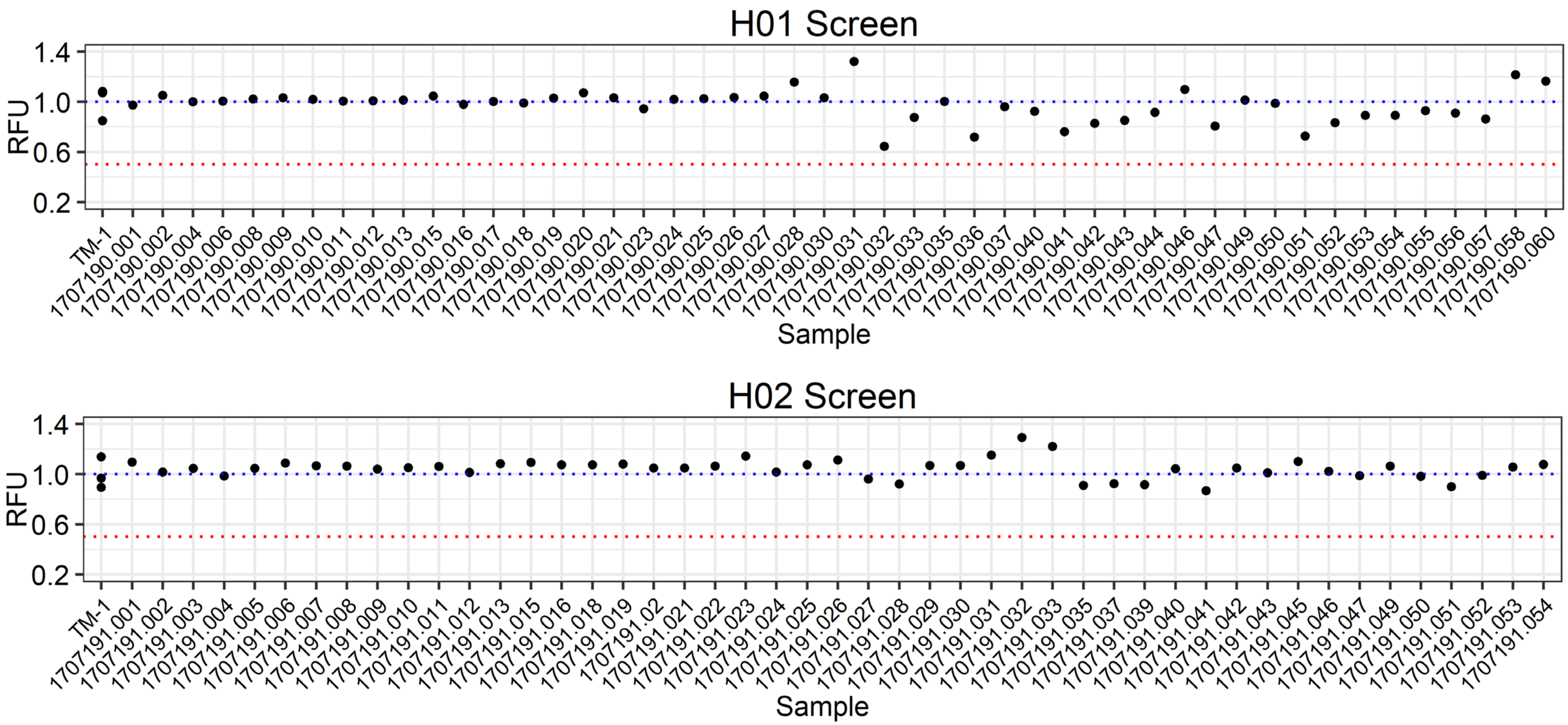

Figure 5.2 Initial screening of putative hypoaneuploids. The identity of the individual sample is shown on the x-axis while the relative fluorescence (RFU) unit normalized to *G. hirsutum* (TM-1) is shown on the y-axis. RFU of "1" indicating euploid value is shown as a blue dashed line while an RFU of "0.5" indicating a monosomic value is shown as a dashed red line.

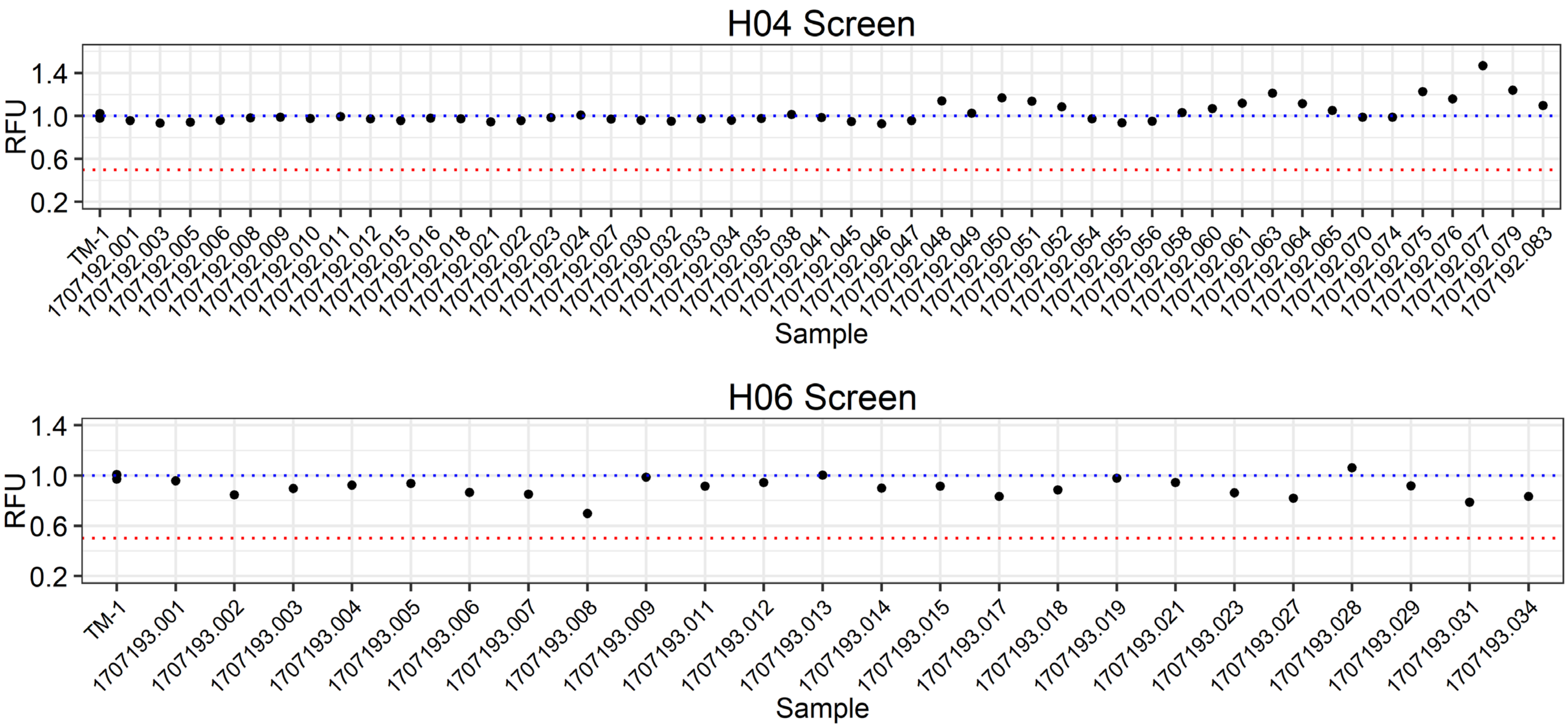

Figure 5.2 (continued)

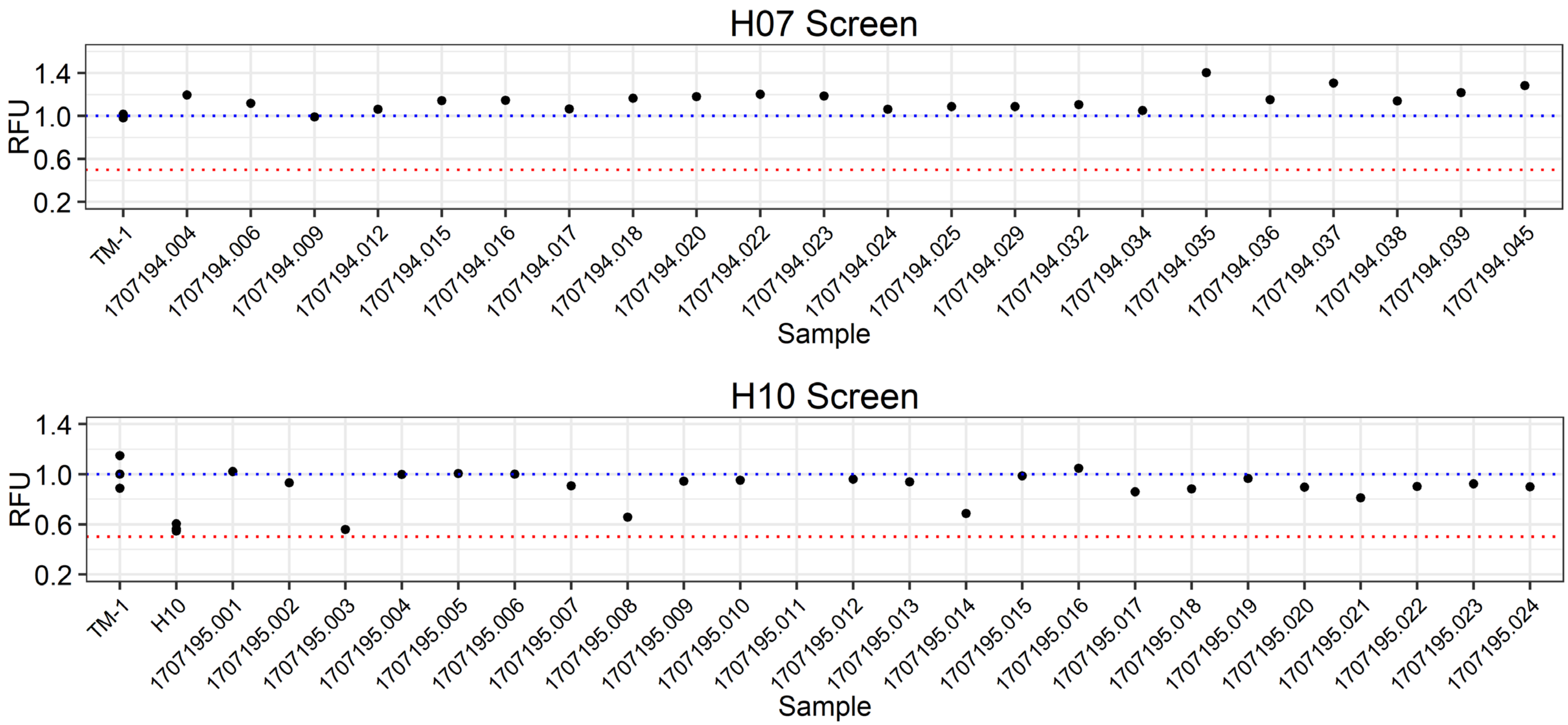

Figure 5.2 (continued)

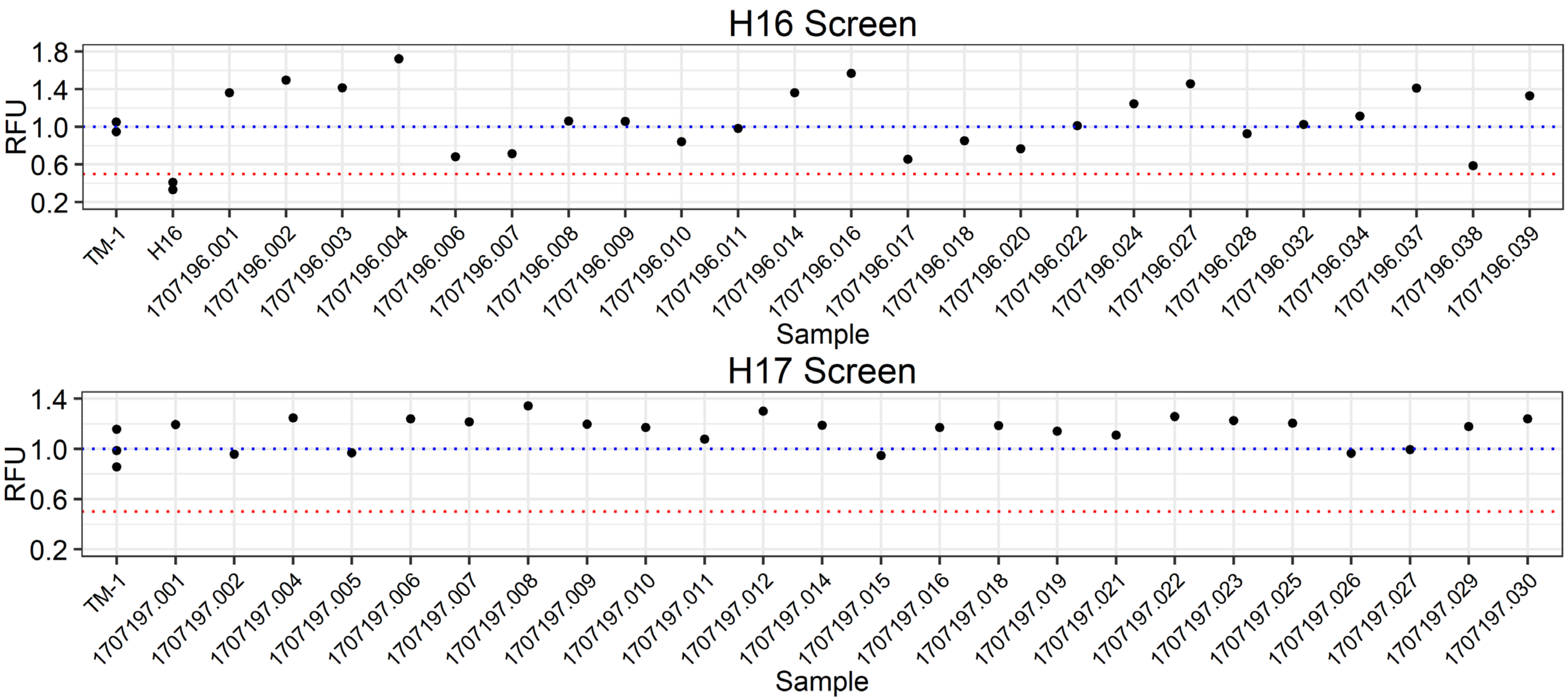

Figure 5.2 (continued)

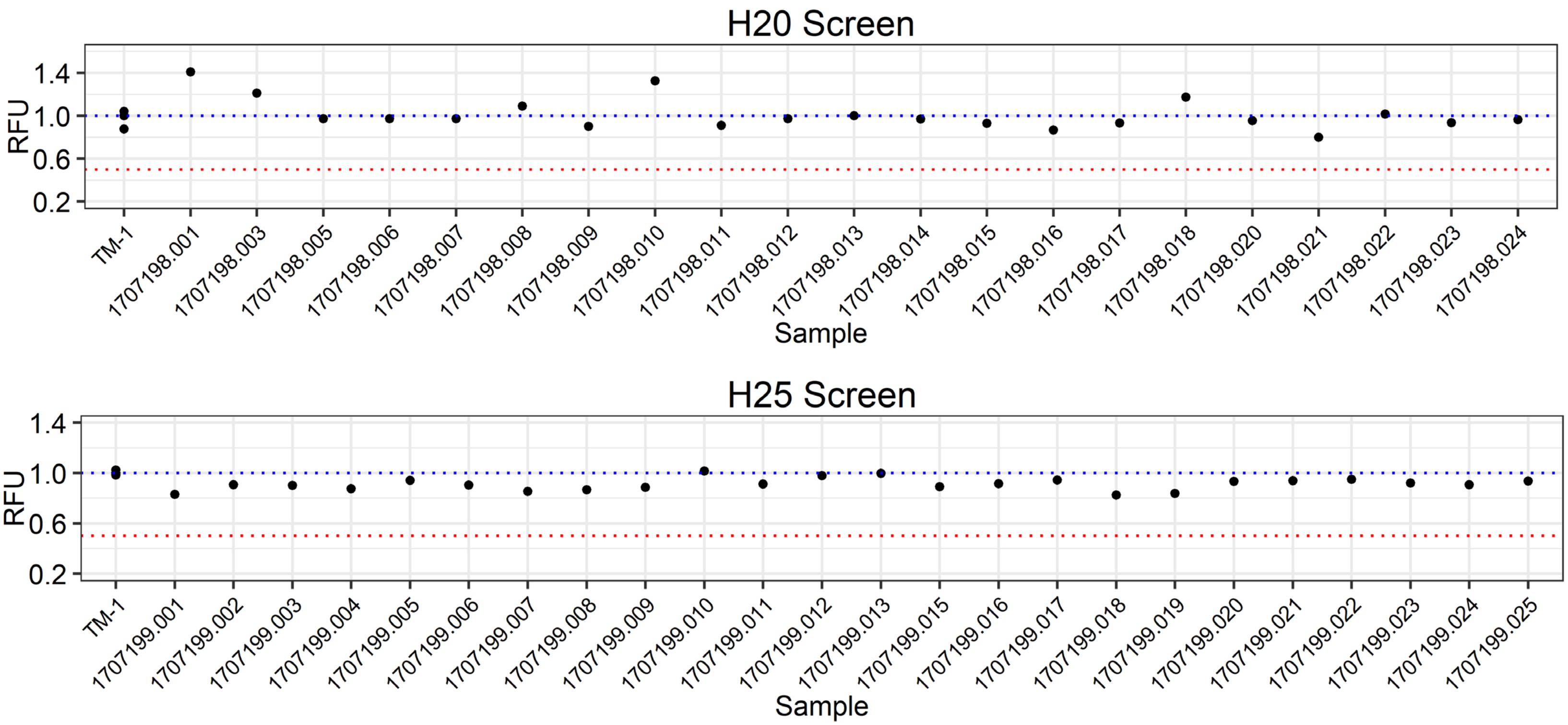

Figure 5.2 (continued)

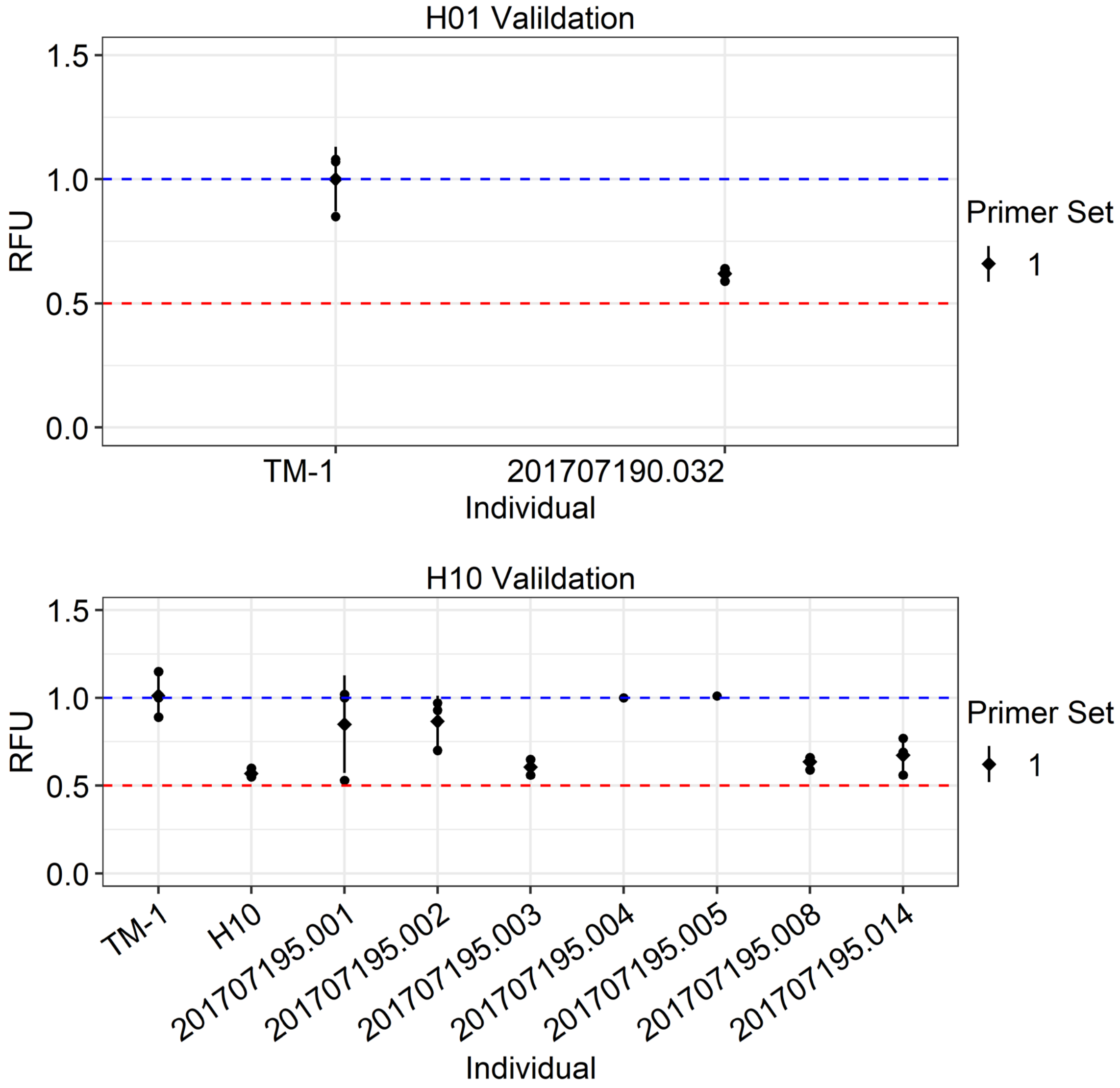

Figure 5.3 Validation screening of putative hypoaneuploids using additional replicates. The identity of the individual sample is shown on the x-axis while the relative fluorescence (RFU) unit normalized to *G. hirsutum* (TM-1) is shown on the y-axis. RFU of "1" indicating euploid value is shown as a blue dashed line while an RFU of "0.5" indicating a monosomic value is shown as a dashed red line.

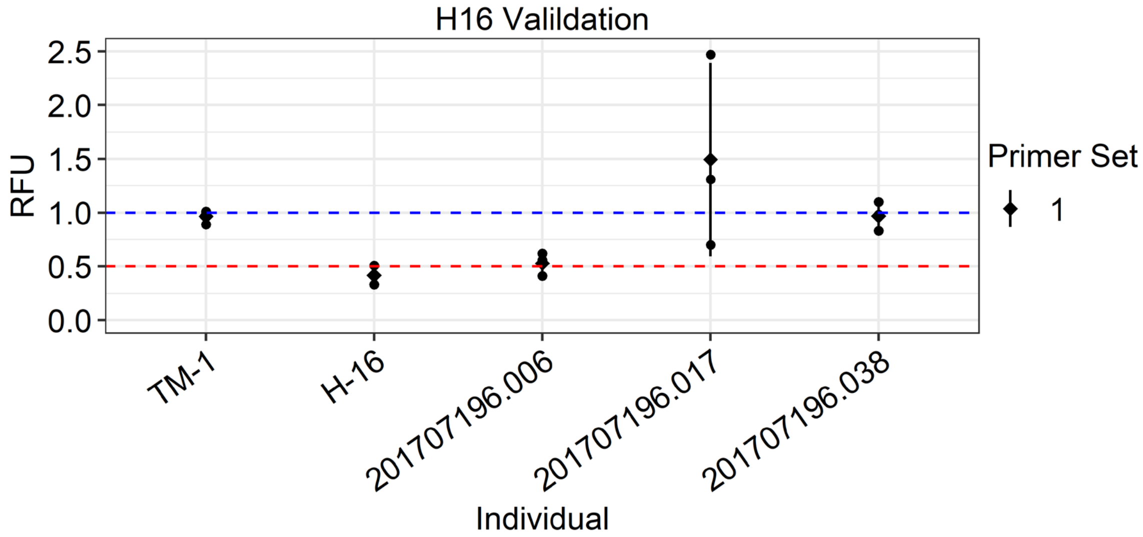

Figure 5.3 (continued)

Development of Monosomic Translocation Heterozygote F_1s

Hypoaneuploids that were successfully identified for being monosomic for their respective chromosome were crossed as females to their respective reciprocal translocation homozygotes to develop the F_1 monosomic translocation

Table 5.1 Germination summary of reciprocal translocation homozygotes.

♂					♀		
Translocation Phenotype	ID	Seed Planted	Germination		Putative Monosome	Monosome Available	F_1
15R-16Ra	8-5Ga	10	Failed	x	H01	Yes	Yes
4R-15L	1040	10	Successful	x	H01		
1L-14L	2780	10	Successful	x	H02	No	No
14L-23	2777	10	Failed	x	H02		
2R-14R	2B-1	10	Failed	x	H02		
20L-22R	DP4	10	Successful	x	H04	No	No
9R-25	2870	10	Failed	x	H06	No	No
20R-25R	2791	10	Failed	x	H06		
1R-16Rb	4672	10	Failed	x	H07	No	No
1R-16Ra	2770	10	Successful	x	H07		
9R-20L	2772	10	Successful	x	H10	Yes	Yes
1L-20R	4669	10	Successful	x	H10		
1L-7L	5-4c	10	Failed	x	H16	Yes	No
7L-12R	1043	10	Successful	x	H16		
7L-18R	4659	10	Failed	x	H16		
1L-3L	2935	10	Failed	x	H17	No	No
2R-3La	IV$_1$	10	Failed	x	H17		
3R-9R	8-30-5	10	Successful	x	H17		
6L-10R	Z9-9	10	Failed	x	H20	Yes	Yes
10R-11R	2785	10	Successful	x	H20		
10R-19R	1626	10	Successful	x	H20		
10L-21L	4675	10	Successful	x	H20		
3L-6L	4010	10	Successful	x	H25	Yes	Yes
6L-7L	1048	10	Failed	x	H25		

heterozygotes (**Table 5.3** and **Figure 5.1A**). The parental lines were crossed over three separate seasons across 2018 and 2019 due to repeated destruction of the F_1 population. The spring of 2018 witnessed a hailstorm that severely damaged the greenhouse in which the parental lines and F_1s were housed in. The summer of 2018 witnessed several heatwaves which resulted in the abortion of F_1 bolls. The extreme heat also facilitated a decrease resistance to fungal pressure within the greenhouse which destroyed surviving bolls. The parental lines were once again crossed during the late Fall of 2018 under overcast, short-day conditions resulting in stunted growth of the F_1 population which was later destroyed at the beginning of 2019 by thrips. The parental lines were successfully crossed during the early spring of 2019, albeit following cotton leaf dwarf

Table 5.2 Germination and screening summary of hypoaneuploids.

Phenotype	Planted	Germinated	Screened	Detected	Recovery (%)
H01	100	72	48	1	2.1
H02	100	85	47	0	0.0
H04	100	58	48	0	0.0
H06	100	79	24	0	0.0
H07	100	63	22	0	0.0
H10	100	77	24	3	12.5
H16	100	77	24	1	4.2
H17	100	78	24	0	0.0
H20	100	95	21	0	0.0
H25	100	90	24	0	0.0

viral infection that stunted their growth, resulting in viable F_1 bolls that were collected early in the summer of the same year.

Screening the Monosomic Translocation Heterozygote F_1s for Hypoaneuploids

The F_1 monosome translocation heterozygote seeds were screened for hypoaneuploidy, as opposed to young leaf tissue, using a high throughput sodium hydroxide based (NaOH) procedure (**Table 5.3**). Screening of the seed was conducted to save time at the cost of decreasing the germination rate due to the mechanical damage that is inflected on the seed. A total of 1,428 F_1 seed were successfully recovered and a total of 987 seed were ultimately screened for hypoaneuploidy using the NaOH procedure **(Figure 5.4)**.This resulted in the 47 individuals with relative fluorescence units (RFUs) that were approximately 40%-60% of their respective euploid controls. These individuals were planted in peat pellets for additional screening utilizing high quality DNA extraction coupled with multiple, unique primer combinations with additional replicates. Out of the five hypoaneuploid families that were originally crossed, only eleven individuals representing four hypoaneuploid phenotypes resulted in detectable transmission of a monosome within the F_1 population (**Figure 5.5** and **Table 5.4**). These were families hypoaneuploid for chromosomes H01, H10, H20, and H25. The highest transmission of a monosome for chromosome H10 with a transmission rate up to 22.22% while the lowest transmission was for chromosome H25 with a transmission rate as low at 0.83% (**Table 5.3**).

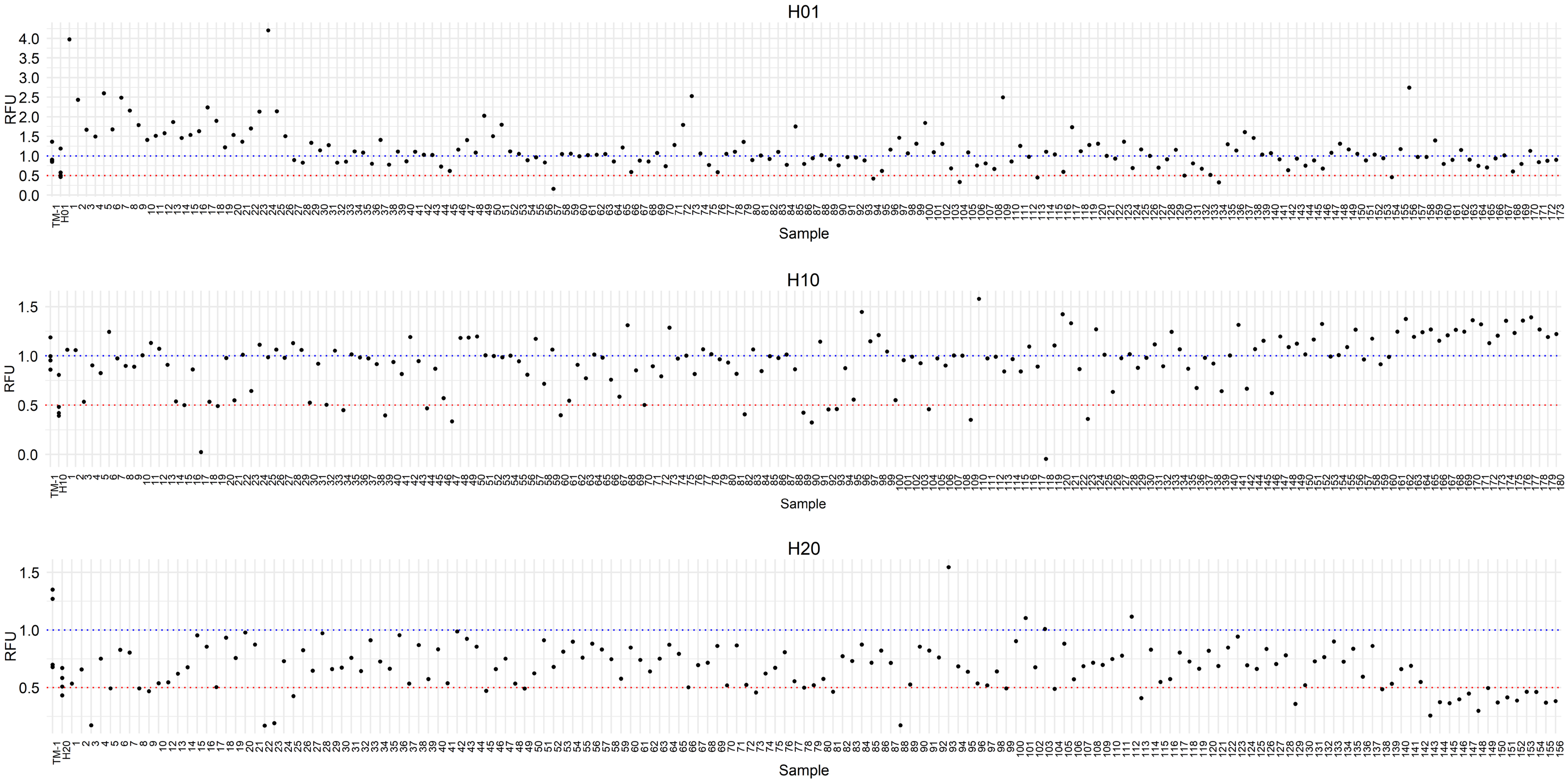

Figure 5.4 Screening of putative F$_1$ hypoaneuploids using high-throughput DNA extraction. The sample number is shown on the x-axis while the relative fluorescence (RFU) unit normalized to *G. hirsutum* (TM-1) is shown on the y-axis. RFU of “1” indicating euploid value is shown as a blue dashed line while an RFU of “0.5” indicating a monosomic value is shown as a dashed red line

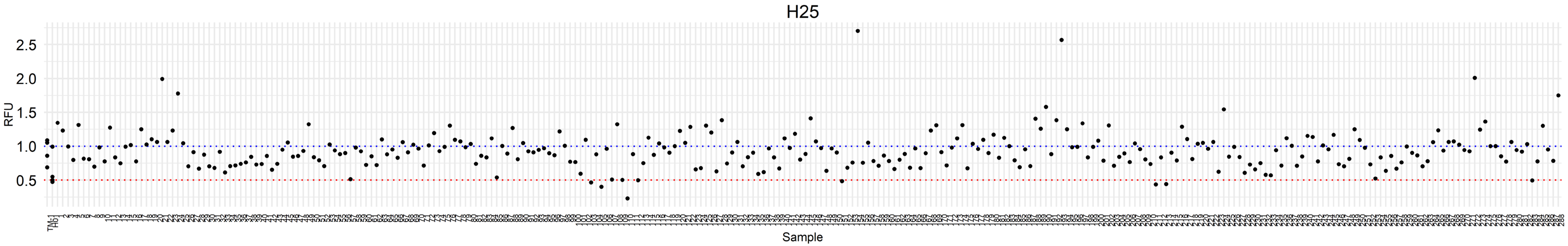

Figure 5.4 (continued)

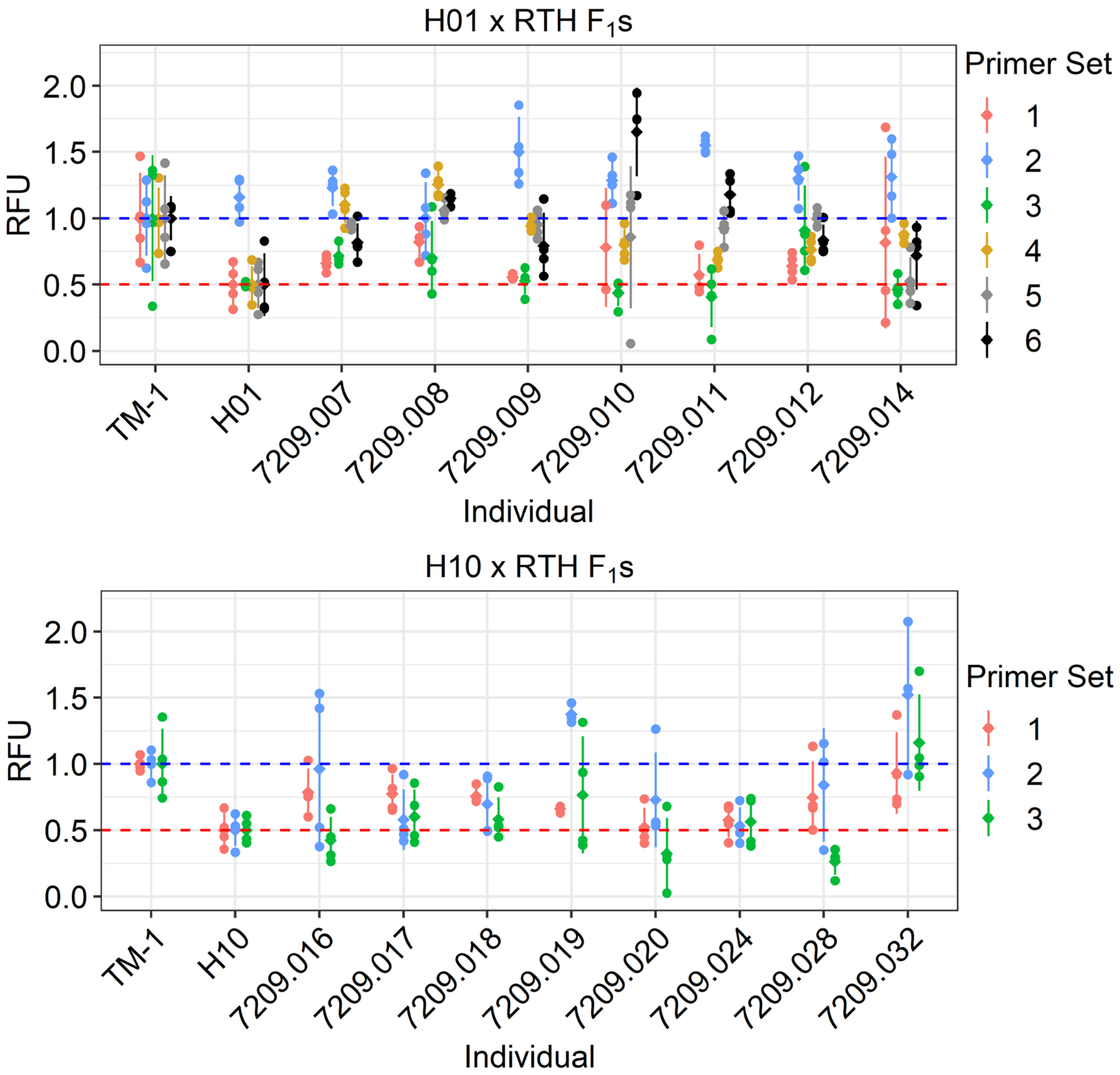

Figure 5.5 Validation screening of putative F_1 hypoaneuploids. The identity of the individual sample is shown on the x-axis while the relative fluorescence (RFU) unit normalized to *G. hirsutum* (TM-1) is shown on the y-axis. RFU of "1" indicating euploid value is shown as a blue dashed line while an RFU of "0.5" indicating a monosomic value is shown as a dashed red line.

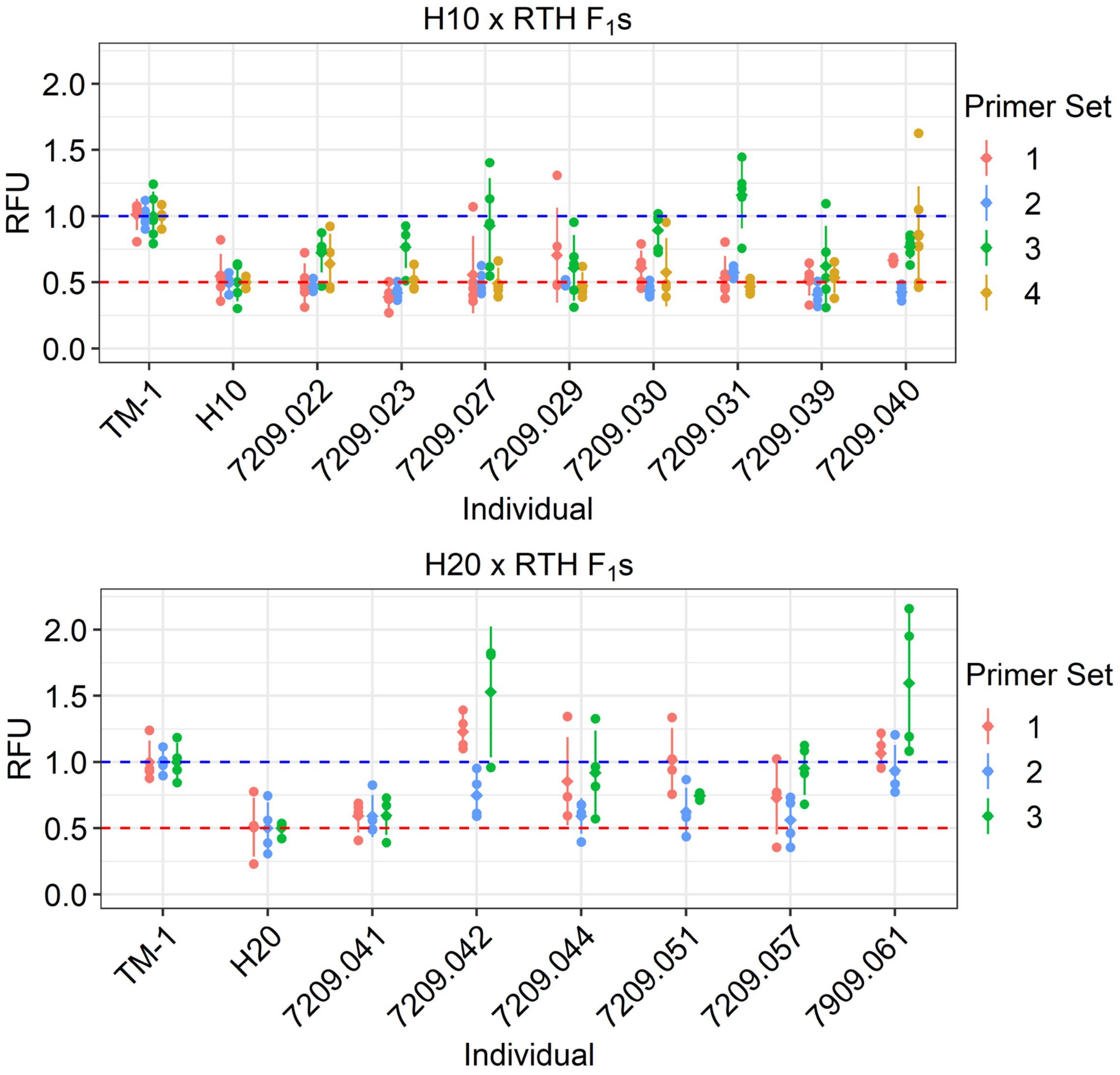

Figure 5.5 (continued)

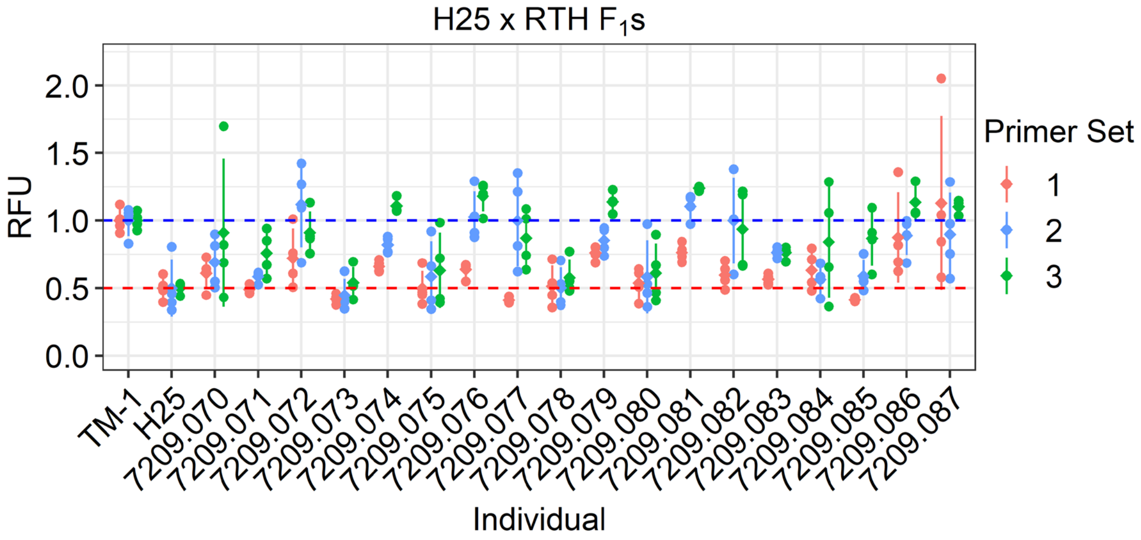

Figure 5.5 (continued)

Table 5.3 Recovery of monosomic translocation heterozygote F_1s.

♂			♀					
Planting #	**Phenotype**		**Planting #**	**Phenotype**	**Seed**	**Screened**	**F_1**	**Recovery (%)**
7190.032	H01	x	7189.01	4R-15L	186	173	3	1.73
7195.003	H10	x	7189.40	9R-20L	11	11	0	
7195.003	H10	x	7189.41	1L-20R	13	12	1	8.33
7195.003	H10	x	7189.42	1L-20R	35	35	0	
7195.008	H10	x	7189.40	9R-20L	15	15	0	
7195.008	H10	x	7189.41	1L-20R	13	9	2	22.22
7195.008	H10	x	7189.42	1L-20R	21	21	1	4.76
7195.014	H10	x	7189.40	9R-20L	48	47	0	
7195.014	H10	x	7189.41	1L-20R	45	41	0	
7195.014	H10	x	7189.42	1L-20R	50	0	0	
1608013.10	H10	x	7189.41	1L-20R	98	85	1	1.18
1608013.10	H10	x	7189.42	1L-20R	14	12	0	
1608014.01	H10	x	7189.41	1L-20R	70	53	0	
1608014.01	H10	x	7189.42	1L-20R	13	11	0	
7196.006	H16	x	7189.49	1L-7L	61	53	0	
7196.006	H16	x	7200.041	7L-12R	254	139	0	
1608008.04	H20	x	7189.68	10L-21L	23	23	1	4.35
1608008.04	H20	x	7189.70	10L-21L	110	96	0	
1608008.04	H20	x	7200.067	10R-11R	8	8	0	
1608008.04	H20	x	7200.069	10R-11R	24	22	0	
1608008.04	H20	x	7200.074	10R-19R	7	5	0	
1608010.03	H25	x	7189.71	3L-6L	45	44	1	2.27
1608010.03	H25	x	7200.77	3L-6L	264	240	1	0.42
				Total	**1428**	**987**	**11**	

Table 5.4 Pedigree summary of monosomic translocation heterozygote F$_1$s.

♀			♂			
Parent	**Phenotype**		**Parent**	**Phenotype**		**F1**
7189.01	4R-15L	x	7190.032	H01	=	7209.009
7189.01	4R-15L	x	7190.032	H01	=	7209.014
7189.41	1L-20R	x	1608013.10	H10	=	7209.024
7189.41	1L-20R	x	7195.008	H10	=	7209.022
7189.41	1L-20R	x	7195.008	H10	=	7209.027
7189.41	1L-20R	x	7195.003	H10	=	7209.023
7189.41	1L-20R	x	1608013.10	H10	=	7209.039
7189.42	1L-20R	x	7195.008	H10	=	7209.031
7189.68	10L-21L	x	1608008.04	H20	=	7209.041
7189.71	3L-6L	x	1608010.03	H25	=	7209.073
7200.77	3L-6L	x	1608010.03	H25	=	7209.078

Discussion

Hypoaneuploids have been used in breeding efforts to facilitate the construction of genetic linkage maps though heterozygous deficiency mapping [142, 143] as well as by aiding in the development of chromosome substitution lines from *G. barbadense*, *G. tomentosum*, and *G. mustelinum* [144-146] which have been used to detect genetic and phenotypic resistance to both root-know nematode and fusarium wilt disease in cotton. The use of *G. hirsutum* translocation lines generated through x-rays, neutron bombardment, and gamma rays have also been used to identify chromosomes as well as to generate recombination maps [140, 147-149]. The use of both hypoaneuploid and translocation stocks were used in this study to develop novel germplasm that could be used to investigate the effects of genetic hemizygosity on meiotic recombination within *G. hirsutum*. The recovery of hypoaneuploids that were monosomic for their respective chromosome varied across lines. Recovery of individuals that were monosomic for chromosome A10 was the highest across all screened hypoaneuploids within screening of the parental lines as well as within screening for the F_1 monosomic translocation heterozygotes (**Table 5.2** and **Table 5.3**). This reflects a high tolerance to the negative consequences incurred by losing an entire chromosome such as gene dosage imbalances that perturb the stoichiometry of regulatory molecules leading to disrupted cellular process [150]. This may be due to chromosome A10 of *G. hirsutum* not harboring as many genes that are essential for regulating such stoichiometric relationships within the cell. There were reduced recovery of hypoaneuploids monosomic for chromosome A10 in certain crosses as well (**Table 5.3)** which may indicate that transmission of a monosome, or successful post zygotic development, may be influenced by external environmental factors such as temperature and biotic stresses. The lowest percent recovery of a hypoaneuploid was for monosomes of chromosome D06 (c25). Coincidently, stable QTLs on chromosome D06 have been associated with a variety of important agronomic traits such fiber strength, fiber length, fiber micronaire, and lint index [151-154]. Additional, chromosome D06 has been associated with significant selective sweeps which have been attributed to domestication of agronomically important traits as discussed earlier (**Figure 2.15-2.17**). The recovery of hypoaneuploids monosomic for chromosome D06 which contain a reciprocal heterozygous translocation between the short arm of chromosome A06 and the short arm of chromosome A03 will allow for examining the effects of perturbing homologous recombination through limiting the sites available for the formation of chiasmata on an agronomically important chromosome.

Developing these F_1 individuals also brings this breeding scheme closer to generating novel germplasm with large structural aberrations that increase the chances of homeologous recombination through genetic hemizygosity between homologous chromosomes. If achieved, reporting of induced homeologous recombination between the A and D subgenomes of *G. hirsutum* will be the first and will allow for examination of the effects of exchanging genetic material between homeologous chromosomes, specifically for chromosomes A01, A10, D06 (c25), and D20 (c20) and their respective homeologs (**Table 5.4**). Individuals that have undergone homeologous recombination can be crossed to a euploid individual to restore the wild type karyotype while maintaining a non-reciprocal exchange of homeologous material. This would result in copy number variation (CNV) between homologous chromosomes. CNV effected genes have been related to important agronomic traits in cotton. A study which deep-sequenced 10 cotton accessions, three founding land races and seven cultivars, discovered 989 genes with copy number variation of which several were found to be associated with fiber quality on chromosomes A01 and D06 through a genome-wide association analyses [155]. Instigating homeologous recombination, particularly between chromosomes A06 and D06, may result in copy number variation within *G. hirsutum* that may elicit genetic mechanisms involved in traits of agronomic importance.

CHAPTER VI

CONCLUSIONS

Greater insight into genome-based constraints on genetic improvement of Upland cotton was sought through detailed analyses of sequence diversity, haplotype structure, haplotype diversity and frequencies, linkage mapping, linkage disequilibrium, and their relationships to various annotated features of newly established reference-grade *Gossypium* AD-genome assemblies. To better characterize Upland cotton diversity at the genomic level, and potentially reveal effects associated with domestication and/or modern breeding, diversity-related features were determined and then contrasted for panels of elite Upland types versus non-elite wild and landrace type accessions of *G. hirsutum*. This also provided some perspectives on the origins of haplotypes, i.e., to determine if long-standing haplotypic structures relate to modern breeding efforts, positions of haplotypic structures were related to experimental recombinational frequencies from controlled crosses between cultivars. To discern whether the recombinational recalcitrance of haplotypic blocks might be overcome by wide crosses, e.g., interspecific crosses, we compared intra- versus inter-specific effects. Lastly, to determine if haplotypic blocks could be disrupted through the recovery of viable recombinants, interspecific backcross breeding project germplasm was examined genome-wide for the presence of such events. As a step toward better understanding and manipulating recombinational behavior in the future, recently compared these recombinationally differentiated regions relative to annotation features available for new genome assemblies for Upland cotton and other AD-genome species.

The haplotype structure of cotton was estimated from analyses of CottonSNP63K-based genotypic data on two panels, one comprising 257 diverse cultivars of domestic and international origin and the other panel comprising 71 "wild" cottons of pre-Columbian origin. The number of variants for each haplotypic block was also determined for each panel. Comparisons were used to identify differences in haplotype structure of improved cottons versus the unimproved types, a goal being to identify genomic regions that have been homogenized in the course of domestication, and SNP-referenced "map" of where diversity does and does not exist in and among "improved" cultivars, and "improved" versus "wild" accessions. For some markers, the use of multiple interspecific and intraspecific mapping populations facilitated the resolution ambiguities in marker alignment resulting from repetitive sequence identities between homeologous chromosomes that arose from the hybridization of two diploid cotton species. The estimation of haplotype structures allowed for the identification of an exceptionally large haplotype block on chromosome A08 within both the improved and wild cotton panels suggesting that the block is of pre-Columbian origin. Assembly-referenced comparisons of intra- and interspecific recombination events in mapping populations showed a close correspondence between a large segment with no/low meiotic recombination in cA08 and the large HB in cA08.

The use of nine genetic mapping populations coupled with a high-quality sequence assembly also allowed for the creation of five interspecific and four intraspecific recombination maps using non-linear LOESS regression estimations. The recombination maps enabled the detection of a rare recombination event within the unusually large haplotype on chromosome A08 that spans approximately 72 Mb. That the detected event occurred in a wide-cross *G. hirsutum* x *G. barbadense* F_2 mapping population suggests that wide-crosses, in particular, might provide a key to disrupting HBs and specifically facilitating diversification of the haplotypic blocks of Upland cotton. Complementary evidence supporting this inference was noted in other wide-crosses and HBs. It seems likely that the same wide-cross approach for HB disruption can likely be extended to *G. barbadense* and the two cultivated A-genome diploids. A localized recombination rate estimation using a 1-Mb non-overlapping sliding window was made possible by using one intraspecific consensus map constructed from three intraspecific populations and three interspecific populations. The localized recombination rates were used to identify areas of the genome that were recombining at elevated and suppressed rates relative their respective genome-wide recombination rate distributions. Genomic features including transposable element density and biological pathways were associated with these regions undergoing increased or decreased rates of recombination and provide further insight into the genomic elements and mechanisms responsible for meiotic recombination in cotton.

In addition to the above-described diversity analysis, haplotype definition, linkage mapping and genome assembly-powered bioinformatic analyses, this doctoral effort included significant advances toward the synthesis of two very different types of novel germplasm, both in pursuit of diversifying Upland cotton and further analysis and manipulation of recombination. One is the development of an interspecific chromosome segment substitution lines, and the other is the creation of a platform for inducing and recovering products of homeologous recombination.

A chromosome segment substitution population consisting of donor germplasm from Hawaiian *G. tomentosum* cotton was repeatedly backcrossed into a *G. hirsutum* background genome to disperse the wild donor genome throughout a genetically improved background. The CSSL population was also partially genotyped through allele-specific PCR, bringing members of the panel one step closer to completion. In addition, genotyping of the panel revealed introgressed *G. tomentosum* chromatin that is recombinant within several large haplotype blocks within improved cotton diversity panel. This provides further evidence that wide-cross germplasm introgression can be used to introduce novel haplotypes into regions under high linkage disequilibrium within improved Upland cotton for exploratory analysis.

The second germplasm development aims to create plants having specific hemizygosity and heterozygosity features that promote the occurrence, detection, and recovery of homeologous recombination events. By rendering one or more chromosomes or segments hemizygous, homologous recombination is precluded, and opportunistic homeologous recombination is encouraged, at least theoretically. The ultimate goal is to recover such recombination products for research and crop improvement. Cross-pollinations were made between monosomic and homozygous translocation Upland cytogenetic stocks as parents to recover monosomic translocation heterozygotes. Four F_1 hybrids were synthesized such that each is hypoaneuploid (monosomic) for one chromosome A01, A10, D06 (c25), or D20 (c10) and heterozygous for a simple reciprocal translocation that involves a chromosomal arm homeologous to the monosome and a non-homologous chromosome. Furthermore, crossing these lines with a euploid *G. hirsutum* will allow for the co-recovery of the monosome and homeologous segmental deficiency, the latter caused by adjacent disjunction of the translocation heterozygote, will result in segregation of one of the translocated arms resulting in hemizygosity. It is theoretically expected that the resulting hemizygosity involves a homeolog of the monosomic chromosome, the propensity for ectopic pairing between homeologous hemizygous segments will be extremely high relative to normal or even to a singly hemizygous segment. If successful, this is a novel approach could be developed into a genome-wide platform for instigating ectopic homologous and homeologous recombination within an allopolyploid genome and will allow for the generation of novel allelic combination as well as increase our understanding of recombination in allopolyploid genomes.

The results of this work advance our understanding of genetic diversity within Upland cotton and related wild *G. hirsutum* accessions: diversity *and* recombinational behavior correlate to the haplotype structure. This immediately suggests specific segments are more important than others for given purposes, e.g., for genetic analysis, introgression, recombination, phenotyping, and breeding. The interspecific CSSL germplasm being generated as part of this overall body of work will likely allow for the genetic diversification of Upland cotton and multiple opportunities to determine if targeted efforts can mitigate or overcome the deleterious effects associated with linkage drag. Future research on the haplotypic blocks and recombination may derive value from the knowledge that certain genomic features correlate with various recombination rates between intraspecific and interspecific mapping populations increases our awareness of potentially significant genomic influences on recombination and the generation of novel combinations of alleles at syntenic loci within a given chromosome pair. In the long-term, this might be extended to homeologs, i.e., if the development of the monosomic translocation heterozygotes leads to experiments whereby homeologous recombination can be induced, making it possible to create truly novel types of cotton. Being able to swap homeologous segments would be of genetic interest and possible agricultural value.

REFERENCES

1. Wendel, J.F. and V.A. Albert, *Phylogenetics of the cotton genus (*Gossypium*) - character-state weighted parsimony analysis of chloroplast-DNA restriction site data and its systematic and biogeographic implications.* Systematic Botany, 1992. **17**(1): p. 115-143.
2. Jiang, C., et al., *Polyploid formation created unique avenues for response to selection in* Gossypium *(cotton).* Proc Natl Acad Sci U S A, 1998. **95**(8): p. 4419-24.
3. Senchina, D.S., et al., *Rate variation among nuclear genes and the age of polyploidy in* Gossypium*.* Mol Biol Evol, 2003. **20**(4): p. 633-43.
4. Cifuentes, M., et al., *Genetic regulation of meiosis in polyploid species: new insights into an old question.* New Phytol, 2010. **186**(1): p. 29-36.
5. Soltis, D.E., P.S. Soltis, and L.H. Rieseberg, *Molecular data and the dynamic nature of polyploidy.* Critical Reviews in Plant Sciences, 2011. **12**(3): p. 243-273.
6. Paterson, A.H., et al., *Repeated polyploidization of* Gossypium *genomes and the evolution of spinnable cotton fibres.* Nature, 2012. **492**(7429): p. 423-7.
7. Blanc, G. and K.H. Wolfe, *Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes.* Plant Cell, 2004. **16**(7): p. 1667-78.
8. Gallagher, J., et al., *A new species of cotton from Wake Atoll,* Gossypium stephensii *(*Malvaceae*).* Systematic Botany, 2017. **42**: p. 115-123.

9. Wendel, J.F., et al., *Evolution and natural history of the cotton genus.* 2009: p. 3-22.

10. Applequist, W.L., R. Cronn, and J.F. Wendel, *Comparative development of fiber in wild and cultivated cotton.* Evolution & Development, 2001. **3**(1): p. 3-17.

11. *United States Department of Agriculture Economic Research Service.*

12. Gunstone, F.D., *Vegetable oils in food techonology: composition, properties and uses*. Vegetable Oils in Food Technology, ed. R.J. Hamilton. 2002: Blackwell.

13. *Texas Department of Agriculture Commissioner SID Miller.*

14. Burke, J.J. *Opportunities for improving cotton's tolerance to high temperature*. in *Beltwide Cotton Conferences*. 2001. Memphis, TN: National Cotton Council of America.

15. Ullah, K., et al., *Impact of temperature on yield and related traits in cotton genotypes.* Journal of Integrative Agriculture, 2016. **15**(3): p. 678-683.

16. Reddy, K.R., H.F. Hodges, and V.R. Reddy, *Temperature effects on cotton fruit retention.* Agronomy Journal, 1992. **84**(1): p. 26-30.

17. Oosterhuis, D.M. *Yield responses to environmental extremes in cotton*. Cotton Research Meeting. 1999. Fayetteville, Arkansas: Proceedings of the 1999 Cotton Research Meeting.

18. Ulloa, M., et al., *Enhancing Upland cotton for drought resilience, productivity, and fiber quality: comparative evaluation and genetic dissection.* Mol Genet Genomics, 2020. **295**(1): p. 155-176.

19. Hao, J.J., M.E. Yang, and R.M. Davis, *Effect of soil inoculum density of* Fusarium oxysporum *f. sp.* vasinfectum *race 4 on disease development in cotton.* Plant Dis, 2009. **93**(12): p. 1324-1328.

20. Isakeit, T., et al. *Identification and management of* Fusarium *wilt race 4*. 2019; Available from: https://www.cottoninc.com/cotton-production/ag-research/plant-pathology/identification-and-management-of-fusarium-wilt-race-4/.

21. Cianchetta, A.N. and R.M. Davis, Fusarium *wilt of cotton: Management strategies.* Crop Protection, 2015. **73**: p. 40-44.

22. Land, C.J., et al., *Cultivar, irrigation, and soil contribution to the enhancement of* Verticillium *wilt disease in cotton.* Crop Protection, 2017. **96**: p. 1-6.

23. Zhao, J., et al., *Quantitative trait locus mapping and candidate gene analysis for* Verticillium *wilt resistance using* Gossypium barbadense *chromosomal segment introgressed line.* Front Plant Sci, 2018. **9**: p. 682.

24. Zhang, J., et al., *Genetic analysis of* Verticillium *wilt resistance in a backcross inbred line population and a meta-analysis of quantitative trait loci for disease resistance in cotton.* BMC Genomics, 2015. **16**: p. 577.

25. Zhang, J. *Improving Upland cotton by introducing desirable genes from Pima cotton*. World Cotton Research Conference. 2007. Lubbock, Texas: World Cotton Research Conference.

26. Huckell, L.W., *Plant remains from the Pinaleño cotton Cache, Arizona.* Journal of Southwestern Anthropology and History, 1993. **59**(2): p. 147-203.

27. Wendel, J.F., C.L. Brubaker, and A.E. Percival, *Genetic diversity in* Gossypium hirsutum *and the origin of Upland cotton.* American Journal of Botany, 1992. **79**(11): p. 1291-1310.

28. Zeng, L., *Broadening the genetic base of Upland cotton in U.S. cultivars – genetic variation for lint yield and fiber quality in germplasm resources.* 2014.

29. Stewert, J.M., *Potential for crop improvement with exotic germplasm and genetic engineering*, in *World Cotton Research Conference*. 1994, University of Arkansas: Brisbane, Australia. p. 313-327.

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

31. Dai, B., et al., *Genomic heterozygosity and hybrid breakdown in cotton (*Gossypium*): different traits, different effects.* BMC Genet, 2016. **17**: p. 58.

32. Hu, Y., et al., Gossypium barbadense *and* Gossypium hirsutum *genomes provide insights into the origin and evolution of allotetraploid cotton.* Nat Genet, 2019. **51**(4): p. 739-748.

33. Ouyang, Y. and Q. Zhang, *Understanding reproductive isolation based on the rice model.* Annu Rev Plant Biol, 2013. **64**: p. 111-35.

34. Edmands, S., S.L. Northrup, and A.S. Hwang, *Maladapted gene complexes within populations of the intertidal copepod* Tigriopus californicus*?* Evolution, 2009. **63**(8): p. 2184-92.

35. Burton, R.S., R.J. Pereira, and F.S. Barreto, *Cytonuclear genomic interactions and hybrid breakdown.* Annual Review of Ecology, Evolution, and Systematics, Vol 44, 2013. **44**(1): p. 281-302.

36. Luteyn, J.L. and P.A. Fryxell, *The Natural History of the Cotton Tribe*. Vol. 32. 1979, Texas A&M University Press: Brittonia.

37. Pickersgill, B., C.H.B. Spencer, and D. de Andrade-Lima, *Wild Cotton in Northeast Brazil.* Biotropica, 1975. **7**(1): p. 42-54.

38. Johansson, A. and U. Gyllensten, *Identification of local selective sweeps in human populations since the exodus from Africa.* Hereditas, 2008. **145**(3): p. 126-137.

39. Zhang, K., et al., *Randomly distributed crossovers may generate block-like patterns of linkage disequilibrium: an act of genetic drift.* Hum Genet, 2003. **113**(1): p. 51-9.

40. Ding, K. and I.J. Kullo, *Methods for the selection of tagging SNPs: a comparison of tagging efficiency and performance.* Eur J Hum Genet, 2007. **15**(2): p. 228-36.

41. Lambing, C., F.C. Franklin, and C.R. Wang, *Understanding and manipulating meiotic recombination in plants.* Plant Physiol, 2017. **173**(3): p. 1530-1542.

42. Liu, S., et al., *Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome.* PLoS Genet, 2009. **5**(11): p. e1000733.

43. Choi, K., et al., Arabidopsis *meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters.* Nat Genet, 2013. **45**(11): p. 1327-36.

44. Chen, Z.J., et al., *Genomic diversifications of five* Gossypium *allopolyploid species and their impact on cotton improvement.* Nat Genet, 2020.

45. Kankel, M.W., et al., Arabidopsis *MET1 cytosine methyltransferase mutants.* Genetics, 2003. **163**(3): p. 1109-22.

46. Jeddeloh, J.A., J. Bender, and E.J. Richards, *The DNA methylation locus DDM1 is required for maintenance of gene silencing in* Arabidopsis*.* Genes & Development, 1998. **12**(11): p. 1714-1725.

47. Kent, T.V., J. Uzunovic, and S.I. Wright, *Coevolution between transposable elements and recombination.* Philos Trans R Soc Lond B Biol Sci, 2017. **372**(1736).

48. Sears, E.R., *An induced mutant with homoeologous pairing in common wheat.* Canadian Journal of Genetics and Cytology, 1978. **19**(4): p. 585-593.

49. Riley, R. and V. Chapman, *Genetic control of the cytologically diploid behaviour of hexaploid wheat.* Nature, 1958. **182**(4637): p. 713-715.

50. Martinez, M., et al., *The synaptic behaviour of* Triticum turgidum *with variable doses of the Ph1 locus.* Theoritical and Applied Genetics, 2001. **105**(5): p. 751-758.

51. Prieto, P., P. Shaw, and G. Moore, *Homologue recognition during meiosis is associated with a change in chromatin conformation.* Nature Cell Biology, 2004. **6**(9): p. 906-908.

52. Tyagi, P., et al., *Genetic diversity and population structure in the US Upland cotton (*Gossypium hirsutum *L.).* Theor Appl Genet, 2014. **127**(2): p. 283-95.

53. Mason, A.S., *Interspecific hyridization for Upland cotton improvement,* in *Polyploidy and Hybridization for Crop Improvement*. CRC Press.

54. Qian, L., et al., *Exploring and harnessing haplotype diversity to improve yield stability in crops.* Front Plant Sci, 2017. **8**: p. 1534.

55. Hill, W.G. and A. Robertson, *Linkage disequilibrium in finite populations.* Theor Appl Genet, 1968. **38**(6): p. 226-31.

56. Lewontin, R.C., *The interaction of selection and linkage.* General Considerations; Heterotic Models. Genetics, 1964. **49**(1): p. 49-67.

57. Kim, S.A. and Y.J. Yoo, *Effects of single nucleotide polymorphism marker density on haplotype block partition.* Genomics Inform, 2016. **14**(4): p. 196-204.

58. Barrett, J.C., et al., *Haploview: analysis and visualization of LD and haplotype maps.* Bioinformatics, 2005. **21**(2): p. 263-5.

59. Wang, N., et al., *Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation.* Am J Hum Genet, 2002. **71**(5): p. 1227-34.

60. Gabriel, S.B., et al., *The structure of haplotype blocks in the human genome.* Science, 2002. **296**(5576): p. 2225-9.

61. Allaby, R.G., *Domestication syndrome in plants*, in *Encyclopedia of Global Archaeology*, C. Smith, Editor. 2014, Springer New York: New York, NY. p. 2182-2184.

62. Tanksley, S.D. and S.R. McCouch, *Seed banks and molecular maps: unlocking genetic potential from the wild.* Science, 1997. **277**(5329): p. 1063-6.

63. Allaby, R.G., R.L. Ware, and L. Kistler, *A re-evaluation of the domestication bottleneck from archaeogenomic evidence.* Evol Appl, 2019. **12**(1): p. 29-37.

64. Hinze, L.L., et al., *Diversity analysis of cotton (*Gossypium hirsutum *L.) germplasm using the CottonSNP63K Array.* BMC Plant Biol, 2017. **17**(1): p. 37.

65. Hulse-Kemp, A.M., et al., *Development of a 63K SNP array for cotton and high-density mapping of intraspecific and interspecific populations of* Gossypium *spp.* G3 (Bethesda), 2015. **5**(6): p. 1187-209.

66. Ooijen, J.W., *JoinMap 4: Software for the calculation of genetic linkage maps in experimental populations*. 2006, Kyazma B.V.: Wageningen.

67. Taylor, J. and D. Butler, *R Package ASMap: Efficient genetic linkage map construction and diagnosis.* Journal of Statistical Software, 2017. **79**(6).

68. Camacho, C., et al., *BLAST+: architecture and applications.* BMC Bioinformatics, 2009. **10**: p. 421.

69. Browning, S.R. and B.L. Browning, *Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering.* Am J Hum Genet, 2007. **81**(5): p. 1084-97.

70. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.

71. Hulse-Kemp, A.M., et al., *BAC-End sequence-based snp mining in allotetraploid cotton (*Gossypium*) utilizing resequencing data, phylogenetic inferences, and perspectives for genetic mapping.* G3 (Bethesda), 2015. **5**(6): p. 1095-105.

72. Andrews, S., *FastQC: a quality control tool for high throughput sequence data*. 2010.

73. Chen, S., et al., *fastp: an ultra-fast all-in-one FASTQ preprocessor.* Bioinformatics, 2018. **34**(17): p. i884-i890.

74. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.

75. Li, H., et al., *The sequence alignment/map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.

76. Li, H., *A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.* Bioinformatics, 2011. **27**(21): p. 2987-93.

77. Danecek, P., et al., *The variant call format and VCFtools.* Bioinformatics, 2011. **27**(15): p. 2156-8.

78. Rezvoy, C., et al., *MareyMap: an R-based tool with graphical interface for estimating recombination rates.* Bioinformatics, 2007. **23**(16): p. 2188-9.

79. Team, R.C., *R: a language and environment for statistical computing.* 2018, R Foundation for Statistical Computing: Vienna, Austria.

80. Wickham, H., *ggplot2: Elegant Graphics for Data Analysis*. 2016: Springer-Verlag New York.

81. Alexander, D.H., J. Novembre, and K. Lange, *Fast model-based estimation of ancestry in unrelated individuals.* Genome Res, 2009. **19**(9): p. 1655-64.

82. Fariello, M.I., et al., *Detecting signatures of selection through haplotype differentiation among hierarchically structured populations.* Genetics, 2013. **193**(3): p. 929-+.

83. Bonhomme, M., et al., *Detecting selection in population trees: the Lewontin and Krakauer test extended.* Genetics, 2010. **186**(1): p. 241-62.

84. LiLin-Yin, *CMplot: Circle Manhattan Plot*. 2020.

85. Morgan, T.H., *Random segregation versus coupling in mendelian inheritance.* Science, 1911. **34**(873): p. 384.

86. Sturtevant, A.H., *The linear arrangement of six sex-linked factors in* Drosophila*, as shown by their mode of association.* Journal of Experimental Zoology, 1913. **14**: p. 43-59.

87. Fang, L., et al., *Genomic insights into divergence and dual domestication of cultivated allotetraploid cottons.* Genome Biol, 2017. **18**(1): p. 33.

88. Wang, M., et al., *Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication.* Nat Genet, 2017. **49**(4): p. 579-587.

89. Goldstein, D.B., *Islands of linkage disequilibrium.* Nature Genetics, 2001. **29**(2): p. 109-111.

90. Deniz, O., J.M. Frost, and M.R. Branco, *Regulation of transposable elements by DNA modifications.* Nat Rev Genet, 2019. **20**(7): p. 417-431.

91. Brennecke, J., et al., *An epigenetic role for maternally inherited piRNAs in transposon silencing.* Science, 2008. **322**(5906): p. 1387-92.

92. Zempleni, J., et al., *Repression of transposable elements by histone biotinylation.* J Nutr, 2009. **139**(12): p. 2389-92.

93. Mueller, W.F., et al., *The silent sway of splicing by synonymous substitutions.* J Biol Chem, 2015. **290**(46): p. 27700-11.

94. Gustafsson, C., S. Govindarajan, and J. Minshull, *Codon bias and heterologous protein expression.* Trends Biotechnol, 2004. **22**(7): p. 346-53.

95. Plotkin, J.B. and G. Kudla, *Synonymous but not the same: the causes and consequences of codon bias.* Nat Rev Genet, 2011. **12**(1): p. 32-42.

96. Koehn, R.K., *Evolutionary genetics and environmental stress.* Trends in Ecology and Evolution, 1991. **6**(9): p. 305-306.

97. Parsons, P.A., *Evolutionary rates: effects of stress upon recombination.* Biological Journal of the Linnean Society, 1988. **35**(1): p. 49-68.

98. Shen, C., et al., *Genome-wide recombination rate variation in a recombination map of cotton.* PLoS One, 2017. **12**(11): p. e0188682.

99. Ulloa, M., et al., *Insights into Upland cotton (*Gossypium hirsutum *l.) genetic recombination based on 3 high-density single-nucleotide polymorphism and a consensus map developed independently with common parents.* Genomics Insights, 2017. **10**: p. 1178631017735104.

100. Cleveland, W.S., *Robust locally weighted regression and smoothing scatterplots.* Journal of the American Statistical Association, 1979. **74**(368): p. 829-836.

101. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-2.

102. Alexa, A. and J. Rahnenfuhrer, *topGO: Enrichment analysis for gene ontology*. 2018. p. R package.

103. Grossmann, S., et al., *Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis.* Bioinformatics, 2007. **23**(22): p. 3024-31.

104. Alexa, A., J. Rahnenfuhrer, and T. Lengauer, *Improved scoring of functional groups from gene expression data by decorrelating GO graph structure.* Bioinformatics, 2006. **22**(13): p. 1600-7.

105. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate - a practical and powerful approach to multiple testing.* Journal of the Royal Statistical Society Series B-Statistical Methodology, 1995. **57**(1): p. 289-300.

106. Dinno, A., *dunn.test: Dunn's test of multiple comparisons using rank sums*. 2017.

107. Stephens, M.A., *EDF Statistics for goodness of fit and some comparisons.* Journal of the American Statistical Association, 1974. **69**(347): p. 730.

108. Wang, Y., et al., *MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity.* Nucleic Acids Res, 2012. **40**(7): p. e49.

109. Opitz, S., G. Kunert, and J. Gershenzon, *Increased terpenoid accumulation in cotton (*Gossypium hirsutum*) foliage is a general wound response.* J Chem Ecol, 2008. **34**(4): p. 508-22.

110. Hastings, P.J., et al., *Mechanisms of change in gene copy number.* Nat Rev Genet, 2009. **10**(8): p. 551-64.

111. Bai, Z., et al., *The impact and origin of copy number variations in the* Oryza *species.* BMC Genomics, 2016. **17**: p. 261.

112. Volker, M., et al., *Copy number variation, chromosome rearrangement, and their association with recombination during avian evolution.* Genome Res, 2010. **20**(4): p. 503-11.

113. Wright, S.I., N. Agrawal, and T.E. Bureau, *Effects of recombination rate and gene density on transposable element distributions in* Arabidopsis thaliana*.* Genome Res, 2003. **13**(8): p. 1897-903.

114. Rizzon, C., et al., *Recombination rate and the distribution of transposable elements in the* Drosophila melanogaster *genome.* Genome Res, 2002. **12**(3): p. 400-7.

115. Jansz, N., *DNA methylation dynamics at transposable elements in mammals.* Essays Biochem, 2019. **63**(6): p. 677-689.

116. Brautigam, K. and Q. Cronk, *DNA methylation and the evolution of developmental complexity in plants.* Front Plant Sci, 2018. **9**: p. 1447.

117. Springer, N.M. and R.J. Schmitz, *Exploiting induced and natural epigenetic variation for crop improvement.* Nat Rev Genet, 2017. **18**(9): p. 563-575.

118. Choi, J.Y. and M.D. Purugganan, *Evolutionary epigenomics of retrotransposon-mediated methylation spreading in rice.* Mol Biol Evol, 2018. **35**(2): p. 365-382.

119. Eichten, S.R., et al., *Spreading of heterochromatin is limited to specific families of maize retrotransposons.* PLoS Genet, 2012. **8**(12): p. e1003127.

120. Yelina, N.E., et al., *DNA methylation epigenetically silences crossover hot spots and controls chromosomal domains of meiotic recombination in* Arabidopsis. Genes Dev, 2015. **29**(20): p. 2183-202.

121. Oliver, K.R., J.A. McComb, and W.K. Greene, *Transposable elements: powerful contributors to angiosperm evolution and diversity.* Genome Biol Evol, 2013. **5**(10): p. 1886-901.

122. Schnable, P.S., et al., *The B73 maize genome: complexity, diversity, and dynamics.* Science, 2009. **326**(5956): p. 1112-5.

123. Sabot, F., et al., *Updating of transposable element annotations from large wheat genomic sequences reveals diverse activities and gene associations.* Mol Genet Genomics, 2005. **274**(2): p. 119-30.

124. Hawkins, J.S., et al., *Differential lineage-specific amplification of transposable elements is responsible for genome size variation in* Gossypium. Genome Res, 2006. **16**(10): p. 1252-61.

125. Slotkin, R.K. and R. Martienssen, *Transposable elements and the epigenetic regulation of the genome.* Nat Rev Genet, 2007. **8**(4): p. 272-85.

126. Daron, J., et al., *Organization and evolution of transposable elements along the bread wheat chromosome 3B.* Genome Biol, 2014. **15**(12): p. 546.

127. Plohl, M., N. Mestrovic, and B. Mravinac, *Centromere identity from the DNA point of view.* Chromosoma, 2014. **123**(4): p. 313-25.

128. Witherspoon, D.J., et al., *Alu repeats increase local recombination rates.* BMC Genomics, 2009. **10**: p. 530.

129. Alabi, O.J., et al., *First report of cotton leafroll dwarf virus infecting Upland cotton (*Gossypium hirsutum*) in Texas.* Plant Disease, 2020. **104**(3): p. 998.

130. Khan, M.K., et al., *Genome wide SSR high density genetic map construction from an interspecific cross of* Gossypium hirsutum *x* Gossypium tomentosum*.* Front Plant Sci, 2016. **7**: p. 436.

131. Hou, M., et al., *Construction of microsatellite-based linkage map and mapping of nectarilessness and hairiness genes in* Gossypium tomentosum*.* J Genet, 2013. **92**(3): p. 445-59.

132. Magwanga, R.O., et al., *GBS mapping and analysis of genes conserved between* Gossypium tomentosum *and* Gossypium hirsutum *cotton cultivars that respond to drought stress at the seedling stage of the BC_2F_2 generation.* Int J Mol Sci, 2018. **19**(6).

133. Dai, B., et al., *Identification and characterization of segregation distortion loci on cotton chromosome 18.* Front Plant Sci, 2016. **7**: p. 2037.

134. Griffiths, S., et al., *Molecular characterization of Ph1 as a major chromosome pairing locus in polyploid wheat.* Nature, 2006. **439**(7077): p. 749-52.

135. El Jack Mursal, I. and J.E. Endrizzi, *A reexamination of the diploidlike meiotic behavior of polyploid cotton.* Theor Appl Genet, 1976. **47**(4): p. 171-8.

136. Chambers, S.R., et al., *The mismatch repair system reduces meiotic homeologous recombination and stimulates recombination-dependent chromosome loss.* Mol Cell Biol, 1996. **16**(11): p. 6110-20.

137. Page, J.T., et al., *DNA Sequence evolution and rare homoeologous conversion in tetraploid cotton.* PLoS Genet, 2016. **12**(5): p. e1006012.

138. Selva, E.M., et al., *Mismatch correction acts as a barrier to homeologous recombination in* Saccharomyces cerevisiae. Genetics Society of America, 1995(139): p. 1175-1188.

139. Zhang, T., et al., *Sequencing of allotetraploid cotton (*Gossypium hirsutum *L. acc. TM-1) provides a resource for fiber improvement.* Nat Biotechnol, 2015. **33**(5): p. 531-7.

140. Menzel, M.Y., K.L. Richmond, and B.J. Dougherty, *A chromosome translocation breakpoint map of the* Gossypium hirsutum *genome.* Journal of Heredity, 1985. **76**(6): p. 406-414.

141. Zheng, X., et al., *Non-destructive high-throughput DNA extraction and genotyping methods for cotton seeds and seedlings.* Biotechniques, 2015. **58**(5): p. 234-43.

142. Park, Y.-H., et al., *Genetic mapping of new cotton fiber loci using EST-derived microsatellites in an interspecific recombinant inbred line cotton population.* Molecular Genetics and Genomics, 2005. **274**(4): p. 428-441.

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

144. Saha, S., et al., *Effect of chromosome substitutions from* Gossypium barbadense *L. 3-79 into* G. hirsutum *L. TM-1 on agronomic and fiber traits.* J. Cotton. Sci., 2004. **8**.

145. Ulloa, M., et al., *Analysis of root-knot nematode and* Fusarium *wilt disease resistance in cotton (*Gossypium *spp.) using chromosome substitution lines from two alien species.* Genetica, 2016. **144**(2): p. 167-79.

146. Saha, S., et al., *Molecular confirmation of* Gossypium hirsutum *chromosome substitution lines.* Euphytica, 2015. **205**(2): p. 459-473.

147. Menzel, M.Y. and M.S. Brown, *The tolerance of* Gossypium hirsutum *for deficiencies and duplications.* The American Naturalist, 1954. **87**(843): p. 407-418.

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

149. Brown, M.S., *Cotton from Bikini; chromosome irregularities found in plants grown from seed exposed to gamma radiation.* J Hered, 1950. **41**(5): p. 115-21.

150. Birchler, J.A. and R.A. Veitia, *The gene balance hypothesis: from classical genetics to modern genomics.* Plant Cell, 2007. **19**(2): p. 395-402.

151. Ijaz, B., et al., *Fiber quality improvement in Upland cotton (*Gossypium hirsutum *l.): quantitative trait loci mapping and marker assisted selection application.* Front Plant Sci, 2019. **10**: p. 1585.

152. Zhang, Z., et al., *High resolution consensus mapping of quantitative trait loci for fiber strength, length and micronaire on chromosome 25 of the Upland Cotton (*Gossypium hirsutum *L.).* PLoS One, 2015. **10**(8): p. e0135430.

153. Li, S.Q., et al., *QTL mapping and genetic effect of chromosome segment substitution lines with excellent fiber quality from* Gossypium hirsutum *x* Gossypium barbadense. Mol Genet Genomics, 2019. **294**(5): p. 1123-1136.

154. Shi, Y., et al., *Dissecting the genetic basis of fiber quality and yield traits in interspecific backcross populations of* Gossypium hirsutum *x* Gossypium barbadense. Mol Genet Genomics, 2019. **294**(6): p. 1385-1402.

155. Fang, L., et al., *Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits.* Nat Genet, 2017. **49**(7): p. 1089-1098.

APPENDIX

SUPPLEMENTAL FILES

Supplemental_table_S1.xlsx List of *G. hirsutum* cotton accessions used in Chapter II.

Supplemental_table_S2.xlsx Haplotype structure data used in Chapter II.

Supplemental_table_S3.xlsx Selective sweep data used in Chapter II.

Supplemental_table_S4.xlsx Recombination rate data used in Chapter III.

Supplemental_table_S5.xlsx Gene enrichment data used in Chapter III.