December 2020

Genetic Characterization of the Pee Dee Cotton Breeding Program

Grant T. Billings
Clemson University, granttbillings@gmail.com

GENETIC CHARACTERIZATION OF THE PEE DEE COTTON BREEDING PROGRAM

A Thesis
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
Plant and Environmental Sciences

by
Grant Billings
December 2020

Accepted by:
Michael A. Jones, Committee Chair
B. Todd Campbell
Sachin Rustgi
William C. Bridges

ABSTRACT

The history of cotton breeding in the southeastern United States is multifaceted and complex. Public and private breeding programs have driven cotton's genetic development over the past two centuries. The Pee Dee breeding program in Florence, South Carolina, has had a substantial role in the development of well-adapted cotton cultivars with improved fiber strength, fiber length, and performance in farmers' fields. Despite the historic importance of the cotton germplasm lines and varieties from the Pee Dee program, little has been done to characterize the population structure and genetic architecture of key traits in this closed breeding program. Here, I first provide an in-depth exploration of the rich history of cotton breeding and genetics over the past century to provide some context for the remainder of this thesis. Then, I discuss the interface of breeding goals, population genetics, and historical implications of a representative sample across 85+ years of cotton breeding in the Pee Dee program. Once the family structure had been evaluated, I applied modern statistical methodology to find gene haplotypes that are associated with improved fiber quality or field performance and attempted to trace the origin of some beneficial alleles. Lastly, I talk about the implications of our work and how it may influence future breeding efforts to utilize the germplasm from this diverse cotton collection.

DEDICATION

I dedicate this thesis to all my friends and family, who supported me throughout the process of my degree.

ACKNOWLEDGEMENTS

I would like to acknowledge the thoughtful feedback I received from each of my committee members throughout the preparation of this thesis:

> Dr. Jones, who first gave me the opportunity to learn about agriculture as a West Florence High School student;
>
> Dr. Campbell, who inspired me to join a plant breeding group during my time at Michigan State University;
>
> Dr. Bridges, who spent hours of his precious time teaching me how to use SAS and think about statistics in a way that makes sense;
>
> and Dr. Rustgi, who has taught me so much about molecular biology and genetics and always pointed me in the right direction when I am confused.

In addition to my four committee members, I would also like to thank Dr. Amanda Hulse-Kemp, who provided invaluable mentorship, advice, and guidance, as well as inspiring me to pursue a career in computational biology.

I would like to thank Natalie Kaiser, who mentored me as an Undergraduate and showed me how you can be a great scientist and also caring, encouraging, and challenging. I also could not forget my high school biology teacher, Mrs. Judy Lee, for sending me to work at Pee Dee REC for the first time and introducing me to biology.

I would also like to acknowledge all my fantastic professors at Clemson, especially Dr. Leigh Anne Clark who taught me to love cytogenetics; Drs. Yu-Bo Wang and Dr. Patrick Gerard, both who taught me to think about Statistics in a new way; and all the others who gave me the skillset to complete my degree.

Lastly, I am gracious to Cotton Incorporated for providing generous funding for my degree program.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Page

CHAPTER ONE

LITERATURE REVIEW

History of Cotton

Production Characteristics, Taxonomy, and Evolution of Gossypium *spp.*

Cotton is cultivated for the elongated epidermal cells or fibers that initiate as extensions of single epidermal cells on the outer integument of the ovule. These fibers (or cells) develop into tough hairs which are made up of over 90% cellulose on a dry weight basis (Fryxell 1963). After senescence, the lint and seeds are harvested from the open fruit, processed, and used to make a variety of textiles and other materials useful for humans. Cultivated cottons belong to the genus *Gossypium* (Family: Malvaceae), with approximately fifty species currently accepted {Wendel, 2015 #27} and a half dozen species important for worldwide economic production and scientific inquiry (Wendel and Albert 1992). More specifically, two tetraploids (*Gossypium hirsutum* L. and *G. barbadense* L.) and two diploids (*G. arboreum* L. and *G. herbaceum* L.) comprise nearly all the cultivated cotton grown today (Gillham et al. 1995).

In the United States, *G. barbadense*, also known as Pima or Extra-Long Staple (ELS) cotton, is grown in California, Arizona, New Mexico, and Texas, and comprises 3% of annual production by weight (Johnson et al. 2018). On the other hand, *G. hirsutum*, or Upland cotton, makes up the balance of cotton production and is grown throughout the Cotton Belt, from the West Coast to Virginia (Meyer 2020). Consistently over the past five years, the Southeast region has planted approximately three million acres of upland cotton, more than the Delta region's approximately two million acres and Southwest region’s seven million acres. India, China, and the United States claim two-thirds of worldwide cotton production, with the majority of remaining bales coming from Pakistan, Bangladesh, Turkey, Vietnam, Brazil, and Australia (Meyer 2020).

Growing, ginning, and manufacturing cotton is a major worldwide economic force, worth more than 120 billion USD annually in the United States alone (NCCA 2011). It is an interesting question indeed wondering how cotton became the ubiquitous material it is today, and especially why and how the tetraploid species have elevated vigor and yield. Wendel and Cronn (2003) note that there is likely a variety of mechanisms that contribute to this observed phenotypic difference, notably the “’buffering’ capacity afforded by duplicated genes [… and] the fixed heterozygosity of their duplicated genomes [across sub-genomes].” This principal is akin to a within-individual hybrid vigor (Crow 1948).

The genome of each diploid cotton species is classified into one of eight genomic groups (A, B, C, D, E, F, G, and K) whereas the tetraploids are all of the AD_n group (Wang et al. 2018). Besides Upland and Pima cottons, there are three other described tetraploid cottons: *G. tomentosum* Nuttall ex Seemann, *G. mustelinum* Miers ex Watt, and *G. darwinii* Watt. The clade including the New World AD-group species arose from a single hybridization event one to two million years ago between two diploid species of distinct continental origin (Wendel 1989). The precise donor species in the polyploidization event are unknown. The extant species *G. arboreum*, donating the cytoplasmic and maternal nuclear “A” genome originating from Africa and Asia, and *G. raimondii*, donating the paternal nuclear “D” genome originating from the Americas, are the most broadly accepted extant descendants today (Wendel 1989). Therefore, the majority of modern-day cottons are paleo-allotetraploids.

Although there is little known about the domestication of the tetraploid cottons, it is accepted that cotton has been used by humans for at least 75 years {Chowdhury, 1971 #104;Splitstoser, 2016 #429} and that there were at least four independent domestication events (Wendel et al. 1989). The lack of a strong archaeological record, which exists for other crops such as maize (Wang et al. 1999) or potato (Brush et al. 1995), has also proven problematic. However, archaeogenomics has shed some light on the evolutionary history of *Gossypium*. In archaeo- or paleo-genomics, ancient DNA (aDNA) samples are extracted from well-preserved historical specimens, sequenced, and then aligned to reference genomes to identify polymorphic loci and genome features (Pont et al. 2019). Palmer et al. (2012) used 454 sequencing to identify transposable elements (TEs) common and different from archaeological and present-day samples of *G. herbaceum* and *G. barbadense*. These data showed how TEs were broadly conserved in *G. barbadense*, while major genomic restructuring occurred in *G. herbaceum* samples over the same time period. The complex evolutionary history of *Gossypium* has led to multiple lines of inquiry in bioinformatics and genetics.

Cultivation of Cotton in South Carolina

Cotton was introduced by immigrants to the United States around 1640 at the latest, with the earliest records of cotton cultivation dating to perhaps as early as 1621 (Smith and Cothren 1999). Cotton cultivation began with Sea Island cotton grown along the coast in the Sea Island region of the Lowcountry in Georgia and South Carolina (Kovacik and Mason 1985) and expanded rapidly on smallholder farms of the region (Chaplin, 1991). Sea Island cotton was distinguished from Upland cottons by long, fine fibers extending from smooth, dark black seeds, as compared to the green fuzzy seeds with shorter, rougher fibers from Upland cottons (Kovacik and Mason 1985). The early agricultural system along the coastline was primarily sustenance farming in small land plots, which focused mainly on the production of indigo and rice. Sea Island cotton was introduced to farmers on Hilton Head Island in 1790, probably coming from the West Indies via the Bahamas (Kovacik and Mason 1985). In this new environment, the natively perennial herbaceous shrub was cultivated as an annual (Stephens 1976).

Before the invention of the modern cotton gin, farmers depended on "naked" seed *G. barbadense* cultivars, whose small black seeds easily separated from the lint using hand-separation or the *churka* (roller) gin (Thomas 1965). The roller gin operates by pulling the fibers with rollers or brushes which easily separated from the hard, dark seeds loosely attached to the fibers of long-staple or Sea Island cottons. This method proved ineffective on short-staple cottons because of the strength with which the seed clung to the fibers. However, larger-scale cotton cultivation did not become commercially viable until the invention of the modern cotton gin by Eli Whitney in 1793 (Chaplin 1991). The cotton gin enabled more facile processing for use in the textile industry by introducing a mechanical method to separate the cotton lint from the seed that was effective on Upland, short-staple cottons (Thomas 1965). The explosion of the cotton industry led to many new economic opportunities in the antebellum South, including the raising, marketing, and processing of cotton fiber and seed (Chaplin 1991).

Sea Island cotton cultivation in South Carolina rapidly declined in 1918 with the invasion of the boll weevil (Harris 1919). The industry saw major changes over the following two centuries from 1800 onwards, including the elimination of chattel slavery, improved access to mechanical implements, and a gradual reduction in coastline production of Sea Island cotton, with the final crop grown on Johns Island in 1956 (Stephens 1976; Kovacik and Mason 1985). As a consequence, green seed Upland cultivars gradually played an increasingly important role in the cotton economy of South Carolina, especially in the inland areas where cotton had not previously been cultivated. Eventually, these changes resulted in the present Upland-dominant system, which took advantage of inland, well-drained soils to produce the hardier *G. hirsutum* cultivars.

Cotton remains an important cash crop in South Carolina, but the geography and composition of cotton cultivars has dramatically changed since the nineteenth and early twentieth centuries. Today, cotton is grown in South Carolina along a c. 250 km strip ranging from the exterior edge of the Sandhills region to c. 50 km from the coast. The primary production areas are the Pee Dee, lower Midlands, and Peach Belt regions along the Georgia border. The top five cotton producing counties (Orangeburg, Calhoun, Williamsburg, Lee, and Hampton), accounted for over 50% of the bales produced in SC in 2018, the most recent year for which county data is available (Wells 2019). The top counties by planted acreage in 2018 were Orangeburg, Darlington, Williamsburg, Calhoun, and Lee.

Early Cotton Breeding

South Carolina effectively had two parallel cotton cultivation systems before the arrival of the boll weevil. The first system was the coastal Sea Island system, which included the cultivation of late maturing *G. barbadense* extra-long-staple cottons and the second was the inland *G. hirsutum* system, which utilized earlier maturing Upland cotton cultivars. Before the introduction of formalized cotton breeding entities, farmers would acquire seed from neighbors or researchers and plant from the same seed stock each year. There was very little phenotypic selection during this era, so hybridization with other cotton strains was commonplace. Most plant selection occurred on the seed level where farmers selected the "best looking" seed to be planted for next year's crop (Moore 1956). During this early time period of U.S. cotton production, only three Upland cotton strains (Georgia Green Seed, Creole Black Seed and Burling's Mexican Hybrid) served as the founding cultigens in the North American Upland cotton gene pool (Calhoun et al. 1997).

In the Mississippi Delta region, Henry W. Vick introduced the concept of single plant selection to cotton breeding in 1939 and used this method to select and reselect superior plants for increased plant vigor in the field. Vick used plant selections to develop the "100 Seed" cotton from Burling's Hybrid seed, which was widely distributed throughout the cotton belt (Moore 1956). His work resulted in the development of pedigreed cottonseed, whose authenticity and parentage were supposedly verified by the producer and seller of the seed. Later work to further improve Upland cotton quality focused on transferring beneficial quality traits from Sea Island cotton into Upland cotton cultivars.

The application of scientific methods to plant breeding gained new prominence in South Carolina when David R. Coker began experimenting with cotton breeding in Hartsville. Coker used new hybridization techniques he learned from his friend Herbert John Webber (a plant physiologist) to develop new methods for efficient breeding of Upland cotton. Coker and Webber identified variability in the plant material in their fields and effectively isolated beneficial crosses of Upland and Sea Island cotton. Coker's work would later be formalized as the Coker Pedigreed Seed company, one of the most influential cotton breeding programs of the twentieth century (Coclanis 1999) and changed the landscape of the Upland cottonseed industry forever (Coclanis 2001).

Cotton Breeding at Pee Dee Research Station

At the same time that Coker was experimenting with breeding improved cotton strains in Hartsville, Florence researchers with the United States Department of Agriculture and South Carolina Agricultural Experiment Station (SCAES) were working on similar issues in parallel (Ware 1937). Since these two breeding programs were located only 30 km apart, there appeared to be significant exchange of germplasm resources between Coker Pedigreed Seed Company and their public counterparts (Calhoun et al. 1997). Breeding at the SCAES/Pee Dee Station began in 1900 when J. S. Newman crossed Upland and Sea Island cottons. In 1911, , H. W. Barre and L. O. Watson collaborated with Orton and Gilbert of USDA Bureau of Plant Industry to identify variability in *G. hirsutum* var. Dixie and others to wilt-resistance. This work was continued by C. A. McLendon until 1920 when all of the breeders left the Florence research station for other work opportunities (Ware 1937).

The breeding program at Florence was restarted and formalized in 1935 as part of the USDA Agricultural Research Service's (ARS) goal to revitalize Sea Island cotton cultivation (Harrell 1974). One of the major impediments in Sea Island cultivation was the preeminent threat posed by the boll weevil, so new breeding strategies were formalized by the station cotton breeders, D. C. Harrell and W. H. Jenkins (Harrell 1974). The Sea Island breeding program was moved to Tifton, GA, via Johns Island, SC, in 1948; however, breeders at the Pee Dee Station took advantage of these genetic resources and their experience with *G. barbadense* to execute intricate and complicated breeding plans (Harrell 1974).

With their breeding objectives now focused on extra-long staple Upland cottons for the Southeast, Harrell and Jenkins examined the crosses and selections from their program. Many of the early crosses used pollen from *G. barbadense* var. Puerto Rico Sea Island and Upland cultivars grown by station agronomist E. E. Hall as the seed parent (Harrell 1974). Thousands of crosses were generated within and between their breeding materials in an effort to combine the fiber quality traits of Sea Island with the agronomic qualities of Upland cotton. A changing focus in breeding goals across the history of the program helps delimit eight specific periods over the history of the program whose germplasm releases reflect those goals (Campbell et al. 2011).

Group one reflected a focus on improved fiber quality by introgressing chromosome segments from *G. barbadense* into reliable, known Upland cultivars, as well as a small focus on a generic-wilt resistance. These releases resulted from the crosses Jenkins and Harrel made in the 1930s and 40s. Group one is particularly important because it represents the selection of major founders in the Pee Dee breeding program population. Parental material in group one came from a variety of places, mainly other cotton breeders, both in the form of varieties and wild accessions.

Breeding efforts during the group one era consisted of complex introductions of alleles from both Upland and non-Upland cottons. Alleles from non-Upland cottons were introduced from *G. barbadense* var. Bleak Hall, a Sea Island cotton previously cultivated in the Lowcountry of South Carolina (Harrell 1974); *G. hirsutum* var. Acala, a putative intercross between Upland and Sea Island cottons in Mexico (Turner 1974); the Triple Hybrid lines, derived from a synthetic tetraploid hybrid *G. arboreum x G. thurberi* crossed to *G. hirsutum* var. Cook 144-133, with multiple backcrossing to *G. hirsutum* var. Coker 100 (Beasley 1940); the experimental *G. barbadense* line 'V' developed by Jenkins; and several unnamed 'Sea Island' and 'Mexican' cottons (Calhoun et al. 1997). Upland alleles were contributed in this cycle by existing elite Upland cultivars and breeding lines, particularly: *G. hirsutum* vars. Coker 100, 100-Wilt and Wilds; California breeding line 'C 6-5' with 'Acala' and 'Hopi Moencopi' background (CAES 1960); 'AHA 6-1-4', an 'Acala' reselection (Culp and Harrell 1973); and various other Upland-allele-dominant cultivars via the products of the Triple Hybrid experiments.

Released germplasm from group one often possessed superior fiber quality, especially increased fiber strength and length, but was also often associated with decreased yield potential. The breeding methods utilized during the group one era involved complex intercrossing, backcrossing, and random mating. Jenkins and Harrell used recurrent phenotypic selection on very large populations to identify favorable recombinants, isolate them, and intermate favorable selections. The gene pool established in group one would play an important role in future developments in the program (Campbell et al. 2011). This material was registered by Culp and Harrell in 1980, although most (if not all) of the crosses were made under the direction of Harrell and Jenkins. It is likely the resources produced in group one were distributed widely before publishing their findings.

Group two had similar goals of improved fiber quality as group one and was largely the "reshuffling" of existing alleles in the breeding program. A major improvement during this breeding era was the introduction of an elite parent, *G. hirsutum* var. Auburn 56 from the Delta Research Station, which conferred resistance to Verticillium wilt (Smith 1964). Group three had a strong focus on improved fiber strength combined with fiber length, including the introduction of alleles from three elite Upland cultivars: *G. hirsutum* vars. Coker 421, Missouri Delta ('MO-DEL'), and Carolina Queen and was the last group to be released by Harrell before he left the program in 1979 (Culp et al. 1979a). Group four broadened the gene pool of the Pee Dee program by introducing 'DSRx6-56', 'Coker 210', and other PD breeding lines from group two. The first new breeding line ('DSRx6-56') was a short, stormproof breeding line developed by Texas Agricultural Experiment Station and was used to increase boll retention and decrease plant height (Culp et al. 1979b). The new line 'Coker 210' was a high-yielding release from the Coker Pedigreed Seed Company (Calhoun et al. 1997). This group included the released varieties 'PD-1', 'PD-2', and 'PD-3' that was intended for use by growers during this period, as well as 'PD875.'

Group five shifted breeding priorities from the improvement of fiber quality to the incorporation of parents with known resistance to insect (Campbell et al. 2011). This generation included the creation of PD695 (a frego bract line) a common parent for the majority of new cultivars in group six. A common donor for insect resistance was *G. hirsutum* var. LA Frego 2, a frego bract line developed at the Louisiana Agricultural Experiment Station. The frego bract (*fg*) trait in cotton was described as a physical marker in the 1950s (Green 1955) and was later associated with resistance to boll rot and the boll weevil. Resistance was conferred by modification of the structure of the bracts surrounding the developing cotton boll, modulating oviposition and decreasing the number of neonates attacking the boll (Jenkins and Parrott 1971). Other cultivars released during group five varied in the presence of the frego bract trait but all displayed some form of insect resistance, indicating that earliness or other cultural changes may have also contributed to resistance (Culp et al. 1990).

Group six included intercrossing between group five cultivar releases in an effort to improve the fiber quality of existing insect resistant lines. The cultivars released in group six are almost all full-siblings or half-sibs, with ‘PD695’ and ‘PD875’ either as one or both parents. The stormproof line ‘DSRx6-56’ was also utilized, as well as frego bract line ‘5-718’ from JB Weaver at Georgia Agricultural Experiment Station and ‘Deltapine 7146N’, a nectariless line with tarnished plant bug resistance. Groups five and six exhibited overall lower fiber quality than previous program releases, likely as a tradeoff for insect resistance.

Group seven involved a continued effort to improve yield in the high fiber quality lines generated at the Pee Dee program. Cultivars from previous breeding groups were crossed with a number of high yielding obsolete cultivars, including ‘Deltapine 41’, ‘McNair 235’ and ‘220’, ‘DES 422’, and ‘Delcot (Delta Cotton) 311’. The PD parents utilized were releases from multiple other breeding groups, especially groups one to four, likely as the donor parents for fiber quality alleles.

Lastly, group eight was focused on improving the presence of desirable recombinants, with recurrent selection upon the cultivars from group seven. A few other sources of genetic variation were introduced, including ‘Coker 315’, ‘Jimian 8’, and a brown lint accession. This group had a breeding goal of breaking the negative linkage between fiber quality and yield.

A survey of American Upland cotton diversity has been undertaken (Tyagi et al. 2014). To date, however, only a single genetic study has been performed to identify patterns of inheritance and genetic diversity within the Pee Dee Germplasm Program specifically, published by Campbell et al. (2009). An additional, thorough phenotypic evaluation (Campbell et al. 2011; Campbell et al. 2012) has provided an invaluable data set that will help inform future genetic endeavors with this closed breeding population. Therefore, in order to more adequately characterize the history of the program and make these resources available for Pee Dee breeders and others, it is important to undertake an in depth genetic survey of the program using newer technology, adding to the volume of resources pertinent to cotton breeders and enabling a future of genomics-assisted decision making.

The Pee Dee cotton breeding program has had at least six breeders. Based on their publication history, I estimate these were the time periods of their tenures: DC Harrell 1935 to 1980; TW Culp 1971 to 1994; CC Green 1990 to 1994; OL May 1995 to 2001; BT Campbell 2004 to present.

Physiology of the Cotton Plant and Fiber

Phenology: Early Growth and Fruiting Initiation

Cotton growth is divided into two overlapping growth stages: vegetative and reproductive. In uncultivated systems, *Gossypium spp.* primarily grow as herbaceous shrubs, usually in a perennial form over most their range (Stephens 1976). The cotton plant grows deep roots during the beginning of its life cycle, providing moderate to strong drought tolerance by tapping into subsurface ground water sources (Ball et al. 1994). The shoot seedling tissue is highly vulnerable to cold temperatures, disease, and mechanical damage. The array of biotic pathogens that attack cotton seedlings is together known as the cotton seedling disease complex (Minton and Garber 1983). *Pythium*, *Fusarium*, and *Rhizoctonia* are three of the most common pathogenic agents responsible for symptoms of the seedling disease complex.

While root development is occurring, the plant diverts energy to increasing shoot leaf area and height. The increase in leaf area over time allows for the plant to generate carbohydrates in excess of that needed for vegetative growth. The exact amount of time required for the vegetative-reproductive conversion to take place is usually described in terms of heat units, or the integral of the temperature-day curve adjusted for a constant temperature component (Reddy et al. 1993). Other factors can control days to flowering, including nutrient availability, water availability, and cultivar selection.

Flowering takes place over the extent of the summer and fall seasons, as long as environmental conditions enable boll retention. It is hypothesized that some amount of stress (i.e., limited nutrient availability) must be present in the environment for cotton to produce the optimal yield, or else the plant may grow prolifically resulting in decreased energy contribution to fruiting (Boquet et al. 1993). However, once the transition to fruiting has been accomplished, it is likely irreversible (Mauney 1966).

Development of the Cotton Fiber

Linters and lint fibers are the two major types of fiber cells that grow from the ovules within a developing cotton boll. Linters are short fibers which adhere to the mature seed during ginning. Cotton lint fibers are longer cells that generally separate during ginning (Stewart 1975). The development of the fiber cells begins at anthesis when the trichome cells differentiate from the outside of the developing seed. Elongation occurs during the first 20 to 25 days, after which primary cell wall biosynthesis ends and secondary wall deposition begins (Gou et al. 2007). The final length of the cells is dependent on a variety of environmental and genetic variables (Paterson et al. 2003). Cellulose fills the secondary cell wall and provides the strong characteristics of the dry fiber bundles (Basra and Malik 1984). The mature bolls crack and fluff open, revealing the mature cotton lint fibers. Therefore, the development of the cotton fibers is dependent on the environmental conditions throughout the c. 50 day development cycle, but overall have strong protection from the environment as long as the boll remains closed (Basra and Malik 1984).

Cotton Fiber Quality Traits

Cotton fiber quality includes the range of physical parameters that describe the characteristics of a sample of mature cotton fibers. Although some cotton fiber parameters may be highly heritable and stable across environments, many are complex traits with major genotype x environment (G x E) interactions and high correlations between traits (Campbell et al. 2012). There are five key fiber traits that generally control the price a grower can receive for a bale of cotton: fiber length, fiber uniformity, fiber strength, micronaire (mic), and color grade (Cotton Incorporated 2017).

Fiber length for *G. hirsutum* is normally in the range 20 mm to 32 mm, whereas extra-long staple cottons (especially *G. barbadense*) ranges from 32 mm to 50 mm. Fiber length is largely cultivar dependent, although nutrient or water limiting conditions can negatively influence fiber length (Jackson and Tilt 1968; Shimshi and Marani 1971). Fiber uniformity describes the ratio of the mean fiber length to the upper-half mean length UHML, where 80% is normal and >85% is highly desirable. Fiber strength, measured in g tex^{-1}, is measured by calculating the force in grams at which a length of 1000 meters of fibers breaks. An average value is 27 g tex^{-1}, whereas a very strong sample will bear >31 g tex^{-1} before breaking. Fiber strength also exhibits high heritability (Campbell et al. 2009). Micronaire is a measure of fiber maturity and fiber fineness. It has a large environmental component by which hot or cold weather and adverse soil moisture conditions can result in overly dense or immature fiber development. Micronaire is preferred in the 3.5 to-4.9 mic units range, and outside of this range the bale loses market value. Color, the final major parameter in fiber quality and valuation, is ostensibly controlled by a few major genes which determine the color type, such as light brown, green, or white (Kohel 1985; Carvalho et al. 2014), although rainfall, storage condition, or ginning can modulate the intensity of the fiber quality for a cultivar (Ware 1932).

The Cotton Genome

Cotton Cytogenetics

The first noted study on cotton cytology was undertaken by Cannon (1903) in collaboration with H. J. Webber of the USDA Bureau of Plant Industry, the same Webber who would later play an instrumental role in the prolific success of Coker Pedigreed Seed Company. Cannon was interested in the formation and fertility of interspecific hybrids, especially in *G. barbadense* x *G. hirsutum*. He used F_1 hybrid seed generated by Webber in South Carolina and planted in a greenhouse in New York Botanical Garden. Cannon observed proper tetrad formation in the gametic cells of the F_1 hybrid microspore cells and had mixed results with respect to self-fertility for the F_1 flowers, indicating that *G. barbadense* x *G. hirsutum* plants had some level of imperfect fertility but indeed appeared to have compatible chromatin (Cannon 1903). W. L. Balls (1910) observed “thread-ring” structure and “black-dots,” in a second later study in Egypt, which had not at the time been reported in meiosis of other plants. Interestingly, Balls and Cannon report different haploid chromosome numbers (n = 20 for Balls, 28 for Cannon), both of which differ from the presently accepted n = 26.

H. J. Denham conducted a survey of *G. barbadense* microspore formation (Denham 1924) and establishment of chromosome numbers for eleven strains of ~7 *Gossypium* species. Denham (1924) correctly established the haploid chromosome number for "American" and "Egyptian" cottons *G. hirsutum* and *G. barbadense* as 26, and the other species in the genus as 13, the presently accepted chromosome numbers for these species Denham posits that perhaps the 26 chromosome species exhibit "gigantism," a hypothesis at the time that identified that plant mutants with double the amount of normal chromatin exhibited vigorous growth.

The next major advancements in cotton cytogenetics came in 1933-1937, when A. Skovsted, a scientist for the Empire Cotton Growing Corporation Cotton Research Station in Trinidad, published his landmark series of four papers "Cytological Studies in Cotton. I-IV."

His first study explored meiosis and mitosis in a sterile triploid hybrid cotton, with 26 chromosomes from *G. arboreum* and the additional 13 from a *G. herbaceum* x *G. arboreum* hybrid (all "A" genome types). He also studied cell division in a fertile *G. herbaceum* x *G. arboreum* cross. Skovsted observed normal cell division and growth in the triploid and diploid hybrids but observed poly-valent formation during prophase I of meiosis. The irregular number of chromosomes observed during metaphase II helped the author identify that the triploid number likely did not represent 3 sets of 13 completely homologous chromosomes. Additionally, reduced chiasma formation in the *G. herbaceum* x *G. arboreum* cross supported this hypothesis. Skovsted also provided two hypotheses explaining the apparently lack of homology between chromosomes in the diploid cottons. They could be polyploids formed by progenitor species of n=6 and 7, or the result of structural rearrangements between chromosomes over time, leading to reduced attraction between chromosomes and therefore irregular n-valent and chiasma formation (Skovsted 1933).

In his second study, Skovsted further explored interesting species hybrids, a fertile hybrid of *G. barbadense* x *G. arboreum* (2n=39) and another infertile complex cross (2n=52) resulting from the aforementioned *G. herbaceum* x *G. arboreum* hybrid crossed with (*G. hirsutum* x *G. barbadense* backcrossed to *G. barbadense*). Skovsted notes that the New World cottons *G. hirsutum* and *G. barbadense* have 13 pairs of small and large chromosomes, of which the small chromosomes from the Asiatic cottons only pair with the small New World cotton chromosomes but the larger chromosomes are always left unpaired. Skovsted correctly made the determination that New World cottons are allotetraploids formed by hybridization between an Asiatic cotton and an unknown other species, with chromosome number doubling occurring at some point (Skovsted 1934). Skovsted established the first strong cytogenetic evidence of the evolutionary origin of the New World cottons with 26 chromosomes.

Lastly, Skovsted demonstrated that chromosomes from Asiatic hybrids paired during metaphase, indicating the presence of homologous chromosomes between the three A-genome diploids tested. He suggested that the American diploids are likely related because of the shared presence of 13 small chromosomes in haploid cells, as opposed to the larger chromosomes in Asiatic species. However, he did not find evidence that the American diploids he examined are likely one of the contributors of genetic content in tetraploid New World cottons (Skovsted 1937). Harland would soon find homologous loci between diploid and tetraploid American cottons, showing the transfer of a "factor" for crinkled leaf and petal spot (Harland 1937).

Recombination Mapping

Recombination mapping, or the use of recombination rate to determine gene position and order, was employed early in cotton genetics. Stephens (1955) used backcrosses to determine the frequency of parental and non-parental phenotype combinations from stable tester stocks. He was able to identify four linkage groups with more than one locus and seven independent loci, a total of eleven linkage groups, far less than the 26 chromosomes in *G. hirsutum.* Stephens also postulated that mutant loci "clustered" in a few chromosomes, suggesting potentially a higher mutation rate or gene density on particular chromosomes. He also speculated that perhaps the tetraploid nature of cotton made the discovery of recessive alleles challenging, since the dominant allele may be present in the other genome (Stephens 1955).

In a review nearly 30 years later, the authors point out that only 61 mutant loci have been pinned to one of 16 linkage groups, of which only 11 had been associated with a particular chromosome (Kohel et al. 1984). The slow process of identifying mutants in the highly redundant tetraploid cottons have posed many problems throughout the process of generating a genetic map for *G. hirsutum* and *G. barbadense*.

Eventually, newer genetic markers took the place of the much slower mutant phenotype-genotype marker system. The first major genetic map for tetraploid cotton was released in 1994, utilizing restriction fragment length polymorphisms (RFLPs) from an inter-specific cross of *G. barbadense* x *G. hirsutum*. Although incomplete, the markers split into 41 linkage groups, of which some mapped to a total of 14 known chromosomes (Reinisch et al. 1994). This work would be further improved by some of the same researchers in 2004, when a near-complete genetic map of tetraploid cotton was released based on the sequence-tagged site (STS) platform. The new map covered over 2,500 loci and served as an important milestone in cotton genetics research (Rong et al. 2004).

Genotyping Technologies

In cotton, two types of genetic markers have been most important for diversity and trait evaluation. Specifically, the PCR-amplified types, such as simple sequence repeat (SSR) or RFLP, and newer, more common single nucleotide polymorphism (SNP) markers. Generally, the PCR-amplified marker types are based on the change in the length of segment of DNA, whereas the SNP markers are called based on the nucleotide base present at a specific genomic position relative to a reference sequence.

The first significant use of genetic markers was in 1980, when the first RFLP map of the human genome was made (Botstein et al. 1980). The early mapping work in man led to future advances, including a high-quality reference genome available today. Likewise, a similar approach was utilized in cotton. However, it was found difficult to identify nucleotide diversity, classically used to describe genetic diversity, in *Gossypium* (Small et al. 1999). The same phenomenon was also observed in terms of nucleotide diversity that would lead to variable SSR genotypes, further complicating the next-best available genotyping method in the 2000s (Rungis et al. 2005). Therefore, newer genotyping methods capable of detecting the so-far undetectable genetic variation were desired.

Current genotyping strategies have been instead based on SNP markers. The array-based methods are currently the least expensive and highly standardized; however, genotyping by sequencing (GBS) may soon overtake array technology. The most widely used and publicly available array platform in the United States is the CottonSNP63K array, developed by Hulse-Kemp et al. (2015). The array is based on the principles of probe-proband hybridization, base extension, and fluorescence, based on the specific fluorescent nucleotide present during the base extension reaction (Gunderson et al. 2005). Competing arrays have also been developed based on a different discovery set of SNPs (Cai et al. 2017).

Finally, genome resequencing presents a different set of challenges and opportunities. Although currently cost prohibitive to detect the low levels of polymorphism relative to repetitive DNA content, lower cost sequencing may make resequencing the most cost-efficient and highly informative genotyping platform (Chen et al. 2007).

Biparental Crosses for Mapping Populations

Biparental crosses have been used for decades in plant genetics to identify large-effect genes underlying desirable traits, or quantitative trait loci (QTL), especially those for resistance to abiotic or biotic stresses, significant changes in plant morphology, quality, and others (Wurschum 2012). Cross-validation in other artificial population is used to authenticate the existence and genomic locations of such high-effect genes.

The approach utilizes two parents which vary significantly for the trait(s) of interest, which are then crossed to generate segregating progeny or progeny families. It is possible to map QTL when the parents have the same phenotypic means if the alleles underlying the phenotype vary (Mauricio 2001). Progenies are then selected for opposing phenotypic extremes, dividing the population into groups of discrete phenotypes. Phenotyping is performed on large numbers of advanced heterozygous lines to increase the probability of observing a large amount of recombination across all chromosomes. Correlation for each marker allele is tested for significant association with the trait of interest, with the lead SNPs carried forward for validation. Composite interval mapping can be used to narrow down the region to specific gene or functional segment.

Biparental mapping populations are helpful when the time and space are available to handle and phenotype large numbers of individuals, especially if linkage disequilibrium decays slowly near the causal gene(s). They are also helpful if natural variation for the trait is difficult to find for formulating a diversity panel. The statistics and experimental design are well established, and this approach has helped to identify many genes or chromosome regions in cotton underlying agronomic or fiber qualities (Xiao et al. 2010; Fang et al. 2010; Fang et al. 2014; Thyssen et al. 2014)

Xiao et al. (2010) used a biparental population by single seed descent to identify the gene underlying resistance to bacterial blight race 18. The resistant cultivar 'Delta Opal' was crossed with susceptible 'DP 388', and 285 families were advanced to the $F_{4:5}$ generation. Phenotyping was performed on families of 21 seedlings in a greenhouse by inoculating cotyledons. Putative homozygous families, or those with all susceptible or resistant plants, had DNA bulked. Simple sequence repeat markers were used to genotype the families and parents. A linkage map from the results was used to identify markers highly correlated with resistance, implicating a subtelomeric gene, *B12*, on chromosome 14. Marker order and linkage was validated in other elite cotton backgrounds. The resulting marker could be used for marker-assisted introgression of the resistance gene.

Thyssen et al. (2014) used next generation sequencing of a near-isogenic line, RNAseq, and bulk segregant analysis to identify the genetic position for the Ligon-lintless-2 (Li_2) locus, a dominant allele underlying a short fiber mutant. Their approach also utilized a biparental population for fine mapping of the locus. The locus had previously been phenotypically assigned to chromosome 18, so the goal of their study was to determine where on the chromosome the causal SNP or polymorphism was located. Near-isogenic line generation was conducted by crossing a short lint mutant with 'DP 5690', then backcrossing to the normal fiber recurrent parent. RNAseq of 8 days post anthesis fibers showed a large density of low lint vs regular lint reads mapping to a telomere of chromosome 13 of the D genome progenitor *G. raimondii* (AD chr18). The F_2 progeny were used for bulk segregant analysis, involving the collection of DNA from phenotypically similar lines, to identify a causal gene based on recombination around the Li_2 locus. Two SNP markers flanked the locus, which contained a single aquaporin gene, although no coding sequence changes were identified in the implicated gene. Nearby transcription factor gene expression varied only for a single C2H2-type zinc finger family protein, with increased expression in the Li_2 mutant 5 DPA. The authors conclude that they could not identify the mechanism by which the aquaporin expression changed, but there was likely a change in a distal control sequence also connected to reactive oxygen species and cellular stress response, affecting cell elongation.

Genome Wide Association Studies

The first genome wide association study (GWAS) was published in 2002, linking disease alleles of the *lymphotoxin-alpha* gene to increased myocardial infarction risk (Ozaki et al. 2002). The principle of such studies is to examine a population of individuals that vary for a trait and then find genetic markers that are highly correlated with that trait. Presently GWAS has expanded across biology to mine for marker-gene associations and serves as an additional tool in the quest to identify causal genes and alleles. In cotton, GWAS has played an important role by serving as a starting point for many studies seeking to identify sources of resistance to disease, fiber improvement, or agronomic performance (Islam et al. 2016; Sun et al. 2017; Huang et al. 2017; Li et al. 2017; Abdelraheem et al. 2020).

Many of the most economically important traits in cotton are complex traits. Meredith and Bridge (1971) identified early on that there were high correlations between many of the complex traits in cotton, suggesting that modified backcrossing or random intermating may be superior for generating favorable recombinants. Campbell et al. (2012) also supported this hypothesis by finding significant correlations between many traits. Therefore, either there are many of the same genes controlling multiple different traits, perhaps negatively in some cases, or different genes are involved with high amounts of pleiotropy. This observation is crucial in planning and understanding GWAS, as the detecting of a positive marker may indeed come at the expense of other important traits, or in fact be caused by an underlying depression in another phenotype. Ingvarsson and Street (2011) further expound on these concerns, especially relevant in terms of complex trait dissection in cotton.

Genome-wide association studies begin with selection of the traits to be studied and a panel of variable individuals. Replicated phenotyping follows, hopefully across a range of environments to evaluate the stability of each genotype's performance. A genotyping platform is selected, and representative samples are processed to make marker allele calls across the genome. A statistical model is identified, like the linear or mixed models, to associate the phenotypic data with each of the genetic marker alleles. During the association step, other information about population structure, stratification, or confounding characteristics may be included as covariates to improve power and decrease the false positive discovery rate. The resulting significant markers are plotted in a Manhattan plot to identify genomic regions with long stretches of associated SNPs. The particular markers can then be used for fine mapping applications to identify candidate genes and undergo validation.

For example, a recent study by Abdelraheem et al. (2020) identified resistance to Verticillium and Fusarium (Race 4) wilts in the US Upland cotton gene pool. A greenhouse study was performed with 367 genotypes, with 4 and 2 complete replications for *Verticillium* and *Fusarium* isolates respectively, whereby infected plants were scored using disease severity ratings; each genotype was replicated twice in a randomize complete block design. The best linear unbiased prediction (BLUP) was calculated for resistance to both resistance traits using a mixed model, treating genotypes as random and other factors fixed. The CottonSNP63K array was selected as the genotyping platform, and polymorphic markers were mapped to chromosomes using BLAST. Population structure was corrected for in the GWAS, performed in TASSEL 5, using a combination of Principal Component Analysis, the K (relationship) and Q (group membership) matrices from STRUCTURE. Putative QTLs were identified by using a sliding 1-15 Mb window (depending on level of confidence). The GWASs were performed separately for each trial of the experiment, and therefore 4 stable QTLs were identified for Verticillium resistance and 2 for Fusarium resistance.

MAGIC Populations

Substantial natural variation exists in wild *Gossypium* species, but there is a paucity of well-described, widely distributed tool sets for breeders to utilize the >50 cotton species in their breeding programs. Breeders can identify sources of variation in germplasm banks or collect accessions in nature, but the return on investment may be so low that such leaps are seldom taken in modern cotton breeding. Thankfully, new sequencing technologies and phenotyping capabilities may make it easier in the future for breeders to utilize sources of variation to the most effective extent possible. One such tool for utilizing genetic variation is the creation of so called MAGIC, or multi-parent advanced inter-crosses, populations. Populations constructed from wild cotton accessions have the capability to introduce the diverse set of alleles breeders desire while reducing the population size necessary to combine most of the alleles (Shim et al. 2018). Li et al. (2016) demonstrated how genetic diversity can be maintained in a cotton

MAGIC population by using 12 founders for yield, insect resistance, and disease resistance.

MAGIC populations have been proven to be effective breeding tools in a variety of crops, such as rice (Bandillo et al. 2013), wheat (Huang et al. 2012), and maize (Dell'Acqua et al. 2015). These populations are generated in four stages (Huang et al. 2015). First, a diverse set of founders are selected based on phenotypic, geographic, or genetic dissimilarity. In mixing, the founders are intercrossed to generate a heterogenous stock or broad genetic base. So-called funnels are then used to mix together parents of diverse origins to generate lines with the background of multiple founders. In maintenance, advanced intercrossing occurs in the second stage, by which lines across funnels are interbred randomly to generate advanced intercross lines, promoting recombination. Inbreeding is then performed. Inbreeding is used to advance advanced intercross lines to a more homozygous state, improving genotyping capabilities especially in polyploid crops. Therefore, the diversity of the original can be utilized and studied without underlying population structure causing problems in genetic analysis (Huang et al. 2015).

The biggest difference between MAGIC and nested association mapping (NAM) lies in how the crosses are generated early on. MAGIC utilized intermating between all or many parents and subsequent shuffling, whereas NAM combines diverse genotypes with a single, well-studied line. Both approaches can be effective for breaking up linkage, but NAM is specifically optimized to use skim-sequencing or GBS to make very large population sizes feasible. NAM has been extremely effective for complex trait dissection in maize since its introduction, and can be especially in other cases where high amounts of diversity exist in the founding parents (Yu et al. 2008).

The original MAGIC population used for genetic studies began development with work by Jenkins et al. (2008), termed the Random Mated Upland Population Cycle 5 (RMUP-C5). Genotypes from the RMUP-C5 and related derivatives were generated by intercrossing 11 elite cotton cultivars in a half diallel design, resulting in 55 families. The families were randomly mated by bulking pollen from all 55 families and pollinating ten flowers for each family. A sample of seed from each cycle of random intermating was collected and selfed for one generation, generating the C_nS_1 populations for each of the 55 families across six cycles. Crossed seed from each *n* intermating generation was used to grow female parents for the C_{n+1} generation. The released material represented the once selfed seed of the sixth-generation intercrosses C_5S_1. Plants from the C_5 generation indicated changes in correlation between multiple fiber qualities and fiber qualities and yield, showing that linkage between causal loci were successfully broken up with respect to the parents.'

The RMUP-C5 served as the important first three steps of MAGIC population development. The fourth step of development, inbreeding, was carried forward by Fang et al. (2014) who selfed C_5S_1 seed for five additional generations to generate C_5S_6 recombinant inbred lines (RILs). To demonstrate the utility of this new breeding resource, they used SSR markers, two years of replicated field data, and the software packaged TASSEL to identify 54 novel fiber quality QTLs. Interestingly, overall allele frequency for each SSR locus in the RILs was highly correlated (r = 0.99) with the allele frequency in the parental generations, with no obvious population structure indicated by STRUCTURE analysis.

A subset of 547 RILs were used by Islam et al. (2016) to functionally characterize a gene underlying fiber quality. Genotyping-by-sequencing was performed to identify 6071 polymorphic SNPs, and 223 SSR markers from a prior study were also included. All genetic markers were mapped to the 'TM-1' reference genome, except 32 SNPs which were not anchored to a chromosome. Phenotypic means were calculated using BLUPs over four environments. The general linear model (GLM) was first used for marker-trait association in TASSEL, with PCA as the covariate. However, the mixed linear model (MLM), with relatedness as an additional covariate, was instead employed due to a high false discovery rate from the quantile-quantile (QQ) plot of p-values. A total of 86 fiber QTLs were identified using these methods. The most promising candidate gene for fiber quality was *GhRBB1_A07*, a gene on chromosome 7 coding for a very large protein, with an 18 bp deletion variant associated with increased quality. Expression analysis and frequency in other elite lines supported their hypothesis. Islam et al. (2016) demonstrate, from start to finish, how the MAGIC population in cotton can be used to identify a candidate gene which can be cross validated in other cotton germplasm, improving the likelihood of success when used in a marker-based selection program.

The next study on the MAGIC population involved whole genome sequencing and fiber quality measurements across twelve location-years, usually in an alpha-lattice design (Thyssen et al. 2019). Reads were mapped to the 'TM-1' reference genome and

SNPs were called with samtools and bcftools. Phenotypic means were calculated with PROC MIXED in SAS for each location-year, with BLUPs calculated for across environment phenotypes. A separate model was used for micronaire to account for year to year environmental differences. Association mapping was performed with the MLM in GAPIT, with 3 eigenvectors of PCA and the kinship matrix as covariates. GWAS was performed for each environment, with 460 SNPs passing the p-value threshold for at least one location-year and trait, with micronaire and upper-half mean length showing the least environmental stability. The QTL of large affect from Islam et al. (2016) was validated, with pleiotropic effects on multiple traits. To remove the effect of the beneficial allele, the analysis was done on the subset without that haplotype, revealing additional significantly associated markers. Their results show how controlling for environmental conditions in the association model, a large number of variants, and an unstructured population can result in robust QTL discovery.

Another 2019 paper from the same group of scientist further focused on dissecting the genetic basis of fiber length, specifically upper half mean length from the same set of field experiments (Naoumkina et al. 2019). The GWAS indicated a non-reference haplotype on chromosome 24, and those RILs carrying the alternative haplotype had shorter fiber length overall. Interestingly, the only parent homozygous for the alternative haplotype had average fiber quality compared to the other parents. Differentially expressed genes (DEGs) from RNAseq exposed 949 differentially expressed genes across a set of four RILs, each representing a combination of the reference/alternative haplotypes and short/long fibers. Gene set enrichment analysis (SEA) implicated genes associated with carbohydrate metabolism, redox, cell wall, secondary metabolism, stress, and transport. Three DEGs were determined to be close to the QTL on chr24, with one protein-kinase superfamily protein having expression difference between long and short fibers RILs, but its association in the trait was ruled out due to observed recombination. The auxin-responsive *GH3* gene was found to be more lowly expressed in longer fiber RILs, reducing the amount of active auxin, and perhaps leading to longer fiber growth.

Lastly, the most recent study based on the MAGIC population evaluated marker trait association for Verticillium wilt resistance (Zhang et al. 2020). Disease severity ratings were given on greenhouse-grown inoculated plants 30d after inoculation, with two soil surface inoculations of spore suspension starting at the two true leaf stage and one week after. Two tests, one per greenhouse, were conducted in a randomized complete block design with two blocks, including ten seedlings for each of the 550 RILs. The two tests were combined with ANOVA as implemented by PROC MIXED in SAS. An interaction between test and genotype for disease severity was detected, as well as a replication effect within each test, indicating confounding environmental or experimental design problems. Polymorphic SNP alleles were identified using Illumina sequencing, which were then mapped to the 'TM-1' reference genome. GWAS was performed with GAPIT to identify QTLs associated with resistance to Verticillium wilt. Only three QTLs were stable between the two tests, with a few candidate genes involved in pathogen response and recognition were implicated.

In conclusion, the MAGIC population and related GWAS studies show an important property of marker-trait association in cotton. Even with a small number of eleven founders, it is still possible to identify substantial amount of variation for an array of traits if there has been enough recombination. These results support the hypothesis that there are many genes and gene combinations that impact fiber quality or agronomic qualities. Here, especially, this is worth noting, since the present study is focused on this very feat. The Pee Dee Germplasm Enhancement Program represents the reshuffling of alleles with substantial recombination and selection; the data from the MAGIC population suggest that, indeed, if causal genes vary within the Pee Dee gene pool, it will be possible to identify these loci, as long as population structure is adequately controlled.

Structural Properties of the Cotton Genome

The first major advancement in *Gossypium* genomics came with the reference genome releases of the tetraploid-progenitor A and D genome diploids, *G. arboreum* (Li et al. 2014) and *G. raimondii* (Wang et al. 2012b), respectively. The c-value for *G. arboreum* and *G. raimondii* correspond to genome sizes of 1,746 Mb and 885 Mb, indicating the A genome is roughly twice the size of the D genome. The difference in genome size is associated with an increase in retroelements (Grover et al. 2007). This also holds true in tetraploid cottons *G. hirsutum* and *G. barbadense* (Hendrix and Stewart 2005; Fang et al. 2017; Wang et al. 2017a). In terms of genome structure, it its notable that there are large syntenic blocks shared between the two genomes, covering c. 80% of the assembled chromosomes. Namely, large rearrangements are observed on chromosomes 2 and 3 of *G. raimondii* relative to the ancestral state in *G. arboreum* and *Theobroma cacao*, and large indels on chromosomes 7 and 8 of *G. arboreum*.

The genome sequences were published soon after for cultivated tetraploid cottons *G. hirsutum* (Li et al. 2015; Zhang et al. 2015) and *G. barbadense* (Liu et al. 2015; Yuan et al. 2015). The *G. hirsutum* genome sequence indicated a larger amount of gene loss overall in comparison to the diploids (relative to the *T. cacao* rooted phylogenetic tree), likely due to gene redundancy, with more gene loss in the A genome than D (Li et al. 2015). Additionally, observations on the ratio of nonsynonymous to synonymous changes indicated an increased amount of positive selection on the A genome and overall a faster rate of evolution (Fang et al. 2017). These changes are thought to have influenced domestication, as well, considering that the D genome progenitor does not have spinnable fibers, indicating greater changes in the A genome may have enabled humans to utilize *G. hirsutum* for fiber production (Wang et al. 2017a). Although there is far less information available about structural variants between *G. hirsutum* and *G. barbadense* compared with that for the diploids, Wang et al. (2017b) report a total of 16 inversions of at least 1 Mb in length, with the largest inversions on chromosomes 11, 12, 14, and 15.

Population-level Analysis

Population Selection and Experimental Design

Population selection and design is critical when performing a genetic experiment. When natural populations are studied, like in ecology studies, a large enough sample size must be collected in order to characterize the population, but relatively small sample sized can be enough to achieve this goal given enough genetic markers (Nazareno et al. 2017). In artificial populations the most important factor is linkage disequilibrium, which can decay slowly or variably, substantially biasing the results of the study (Hamblin et al.

2011). Regardless, selection of the population for study is a critical step in designing a population genomics experiments. Simulation or resampling can be used to test for the minimum number of markers or individuals required to detect a particular level of population structure or separation.

Wang et al. (2012a) showed population size and balance can affect QTL discovery in barley. They show that with high LD and low effect alleles, coupled with unstable traits, a large number of unrelated individuals are necessary in order to detect small effect QTLs. Their results also show how using a combination of STRUCTURE-based membership coefficient (Q) and kinship best enable correction for population structure and robust QTL discovery. Balance was achieved by subsampling the existing barley cultivars to include an equal number of each across the cultivar categories, allowing the overall effect to balance out across the experiment.

Another consideration with experimental design is the phenotyping method used. In agronomic experiments, many traits are influenced by environmental factors, which generate noise when trying to identify small differences in the target parameter between genotypes. Replication across environments can help tease out these effects, which can then be used to calculate more accurate true estimates of the cultivar or allele effect on a trait. Genotype x environment (G x E) effects are particularly important, especially in cotton where they can drive strong influence on many traits (Campbell et al. 2012). Multiple plot replicates and field replicates, when feasible, can increase power and confidence in the results of field experiments.

Statistical Methods for Population Structure Analysis

A single marker locus describes a single place in the genome, located at a specific physical position on a chromosome. If the physical position is not known, a genetic or map position can be used instead. The genetic marker can be any polymorphism previously described. Functional regions around a genetic marker can affect how a marker is used. For example, if a marker allele is in high LD with a negative allele, a breeder may select against the presence of the negative allele, reducing the frequency of the marker allele in the population. To determine if a marker locus is appropriate for use in a study, filtering parameters are first applied to maintain a minimum level of quality. The marker can be described by its frequency in the population, given as the minor allele frequency (MAF) for SNP markers, calculated by determining the overall proportion of the rarer allele in the population. Marker loci are typically excluded if the MAF < 0.10, 0.05, or 0.025, depending on the application. Call rate (CR) is the proportion of individuals successfully genotyped. A CR < 0.90 or 0.75 are typical, depending on the genotyping platform and application. Whether or not an allele call is missing completely at random (MCAR) or not plays an important part in the CR selected; for example, the systematic absence of a marker allele call could indicate a structural variant, which would violate the MCAR assumption. A chi-square test can be used to see if marker allele frequencies violate the Hardy-Weinberg equilibrium (HWE) assumption, which is given loose constraints in breeding systems. HWE is typically used to eliminate genotyping errors (Hosking et al. 2004).

The relationship between marker loci can also be described. The overall distribution of markers across a chromosome is determined by finding the number of polymorphic markers per a given map or physical distance, or across the entire genome. Genomic regions without detected variability can also be identified. Imputation is used to fill in gaps given observed frequencies of recombination (Halperin and Stephan 2009). The square of the correlation coefficient (r^2) between two marker alleles is used to generate linkage maps, identify haplotypes, and remove correlated markers ($r^2 > 0.8$). *D'* can be used instead of r^2 to describe linkage disequilibrium (VanLiere and Rosenberg 2008). These tools together are used to determine if adequate coverage is available for the desired applications.

After characterizing all the marker loci available in a data set, the genetic markers passing filtering can be used to characterize the population composition, given preassigned groups. Wright's F_{ST} is used to estimate genetic separation between groups, with a value close to zero indicating low separation and closer to one indicating higher separation (Wright 1965; Nei 1973). Other methods are used to test group membership or identify the contribution of putative ancestral populations to the observed substructure; STRUCTURE and fastSTRUCTURE are two implementations of this methodology (Pritchard et al. 2000; Raj et al. 2014). STRUCTURE provides the user with a membership coefficient *Q*, which corresponds to the proportion of alleles in an individual that are assigned to a particular ancestral group according to the model. The best number of groups (*k*) is chosen by finding the point on the first derivative of the model fit by *k* graph where the rate of improved fit no longer increases, similar to a maximum likelihood estimate. This method prevents overfitting, since an increased number of groups will always fit as well or better as a smaller number.

Groupings can also be identified *de novo* using clustering techniques, especially on genetic distance calculations. Clustering is best known in genetics for phylogenetic tress, but these analyses may not capture the breeding history accurately in an artificial population. Two other grouping methods exist with different goals, both performed on individuals and their SNP calls: principal component analysis (PCA) and discriminant analysis of principal components (DAPC). PCA optimizes the model fit relative to the data for individuals, whereas DAPC works on the preassigned groups. Therefore, DAPC may be better when looking for "hidden elements" in genetics data where the true goal is to capture the between group variation as much as possible, rather than between individual variability (Jombart et al. 2010).

Genotype x Environment Dissection in GWAS

Phenotypic stability is of paramount importance for plant breeders because environmental conditions for crop production change year to year, crop cultivars are often produced in a variety of locations, and management practices may differ by the end consumer or use case. Therefore, when identifying genes and loci underlying phenotypes, it is also critical to study how the effects are modulated by non-genic factors. Experiments normally attempt to minimize differences due to environmental effects but including such effects in an association model can provide crucial information, especially for breeding applications. Here, I will describe what G x E is, techniques and examples of how geneticist explore G x E interactions, and applications specifically for cotton.

Fisher and Hogben independently introduced the idea of G x E, aimed towards two different circumstances. Fisher introduced the biometric concept, explaining how a treatment could result in different observations if the environmental conditions varied, whereas Hogben introduced the developmental concept, primarily focused on how the development of an organism could be changed if the surrounding environment differed (Tabery 2008). Ideally, G x E is observed as “crossing” in a reaction norm plot, where the rank order of a phenotype for levels of one factor, genotype, changes with different location, rainfall, temperature, or some other environmental condition. Non-crossover G x E interactions are more frequently observed in cotton, wherein the magnitude of an effect changes across environments but the overall rank order does not (Campbell and Myers 2015). The breeding interpretation for this result might be that generally speaking a variety or breeding line that performs well in one area will likely perform well in other areas, compared to competing varieties; however, the magnitude of the difference in phenotypic means between varieties changes across environments.

More generally, G x E describes how different genotypes change in a non-uniform way to a change in environment, an observation which is frequently interpreted as antagonistic pleiotropy, or the opposite additive effect of an allele (Des Marais et al. 2013). In fact, Des Marais et al. (2013) approximated through a literature survey that at least 60% of QTLs in plants exhibit G x E through antagonistic pleiotropy or environment-specific effects. Most detected QTL x E interactions were simply a change in the overall effect of an allele, not that opposite effects across environments. These changes in general are frequently referred to as phenotypic plasticity in the literature.

Genotype x environment interactions can be observed at the cultivar or locus level. Dia et al. (2018) examined G x E for yield in 22 pickling cucumbers varieties over three years at seven locations. Various environmental corrections were considered, such as rainfall or humidity. They used two methods to test for the presence of G x E interactions. Stability analysis was used, where performance for genotypes at each location is normalized by an environmental index, which is the average of all genotypes for each environment. Non-stable genotypes are those that deviate differently relative to the environmental indices, resulting in a sort of environmental effect in the model. The genotype + genotype x environment interaction (GGE) biplot, used in conjunction with PCA, was also used. For all traits studied, significant effects from environment, genotype, and G x E were detected, with crossing effects observed as well. Methods such as those by Dia et al. are useful for initially determining if G x E exists for a trait of interest but cannot alone show whether or not an underlying genetic component itself is impacted by the environment.

van Eeuwijk et al. (2010) provide an excellent overview on QTL discovery and analysis in plants across environments, especially in terms of modelling QTL main and interaction effects. At the macro-scale traits are examined in terms of G x E effects overall for an individual’s genotype, but these effects can also be further separated at the locus level to determine QTL x E interactions. The level of interdependence in the model between loci, as well as the predicted type of interaction, both impact the interpretation of results.

Importantly, there exist models for combining and utilizing data from multiple environments, accounting for the effect of multiple QTL (and their interactions). These methods are based on the whittling-down of a complex model with thousands of putative QTL to one with only the ones that contribute to a large amount of phenotypic variance. One approach is to identify significantly associated QTL for the phenotype, then add the effect of each QTL to the model until an additional QTL falls below the significance threshold. Then, QTL x E interactions can be tested. Another approach is the multi-trait mixed linear model (MTMM), which uses the covariance matrix between two traits to identify the "pooled" effect of two markers on two phenotypes considered together (Korte et al. 2012). The Bayesian multi-trait and multi-environment model (BMTME) is another alternative that incorporates observations across environments by successive estimation and re-estimation of a very large number of model parameters (Montesinos-Lopez et al. 2016; Montesinos-Lopez et al. 2018).

Additional traits can be added to the model to explore pleiotropic effects from QTL, correlations between traits, and G x E. Malosetti et al. (2008) examined five traits across eight environments in an F_2 population of maize. Their work was based on the aforementioned strategy of searching for genome-wide significant markers and then examining those QTLs in more details. Multiple overlapping QTL peaks were found, indicating either pleiotropic or linked loci. The most effective model tested used the direct product of the trait and environment matrices for modelling the genetic covariance. This mixed model proved more effective than treating each trait-environment combination as its own parameter to estimate, decreasing the computational workload.

Their essential finding was that combining traits and environments in a particular statistical way can increase power for QTL detection while also preventing spurious associations from appearing by searching for too many different associations and over-fitting the model.

Another QTL x E detection technique based on Bayesian statistics was demonstrated in Barley by Zhao and Xu (2012). QTL x E interaction are defined as the variance of the estimated QTL effects across environments, which is helpful because it makes logical sense with the idea of what a QTL x E interaction is in the most general sense: a QTL which has variable impact on a trait in different environments. The results are interpreted over the physical course of the genome by overlapping the location of QTL main effects and the QTL x E interactions. In fact, QTL x E interactions could be detected even in genomic regions where main effects are masked, as interaction terms can make true differences from the main effects approach zero. For all eight traits tested, QTL x E interactions were detected, indicating the existence of unstable QTL underlying each trait. Zhao and Xu demonstrate yet another method for testing for QTL stability, using the variance term for QTL effect as the phenotype parameter.

The first major QTL x E analysis in cotton, by Paterson et al. (2003), tested for the interaction by separate QTL detection across years and levels of irrigation. Six QTL for fiber length were detected, of which four indicated QTL x E interaction via differential QTL detection; seven main effect and five interactions for length uniformity; nine and five for elongation; and 25 main effect and 18 interactions for fineness. Their results, although based on a small number of genetic markers, showed what would later become very clear in the genetic dissection of cotton fiber quality traits -- a QTL in one environment would not necessarily show up when tested in a different environment.

Campbell and Jones (2005) used the additive main effects and multiplicative model (AMMI) to detect G x E in South Carolina cotton variety trials. Lint yield and fiber strength showed the largest G x E of all traits tested, with differences in lint yield only showing up in the non-yield-limiting conditions. Non-rank changing interactions were by far the most common. Campbell et al. (2012) tested Pee Dee germplasm to examine G x E across the Southeastern US cotton cultivation region. They used regression techniques similar to those described by Dia et al. (2018). Environmental stability for cotton fiber quality traits varied significantly by breeding group and genotype, with micronaire showing the smallest proportion of variance explained by G x E and length the highest. Their results suggest that QTL x E likely underlies the genetic architecture of fiber quality traits in the Pee Dee germplasm and deserves further evaluation. This has important implications for GWAS in the Pee Dee material, and methods specifically developed for identifying QTL x E in this structured breeding population will need to be developed.

Literature Cited

Abdelraheem, A., H. Elassbli, Y. Zhu, V. Kuraparthy, L. Hinze *et al.*, 2020 A genome-wide association study uncovers consistent quantitative trait loci for resistance to Verticillium wilt and Fusarium wilt race 4 in the US Upland cotton. *Theoretical and Applied Genetics* **133** (2):563-577.

Ball, R. A., D. M. Oosterhuis, and A. Mauromoustakos, 1994 Growth Dynamics of the Cotton Plant during Water-Deficit Stress. *Agronomy Journal* **86** (5).

Balls, W. L., 1910 The Mechanism of Nuclear Division. *Annals of Botany* **os-24** (4):653-665.

Bandillo, N., C. Raghavan, P. A. Muyco, M. A. Sevilla, I. T. Lobina *et al.*, 2013 Multi-parent advanced generation inter-cross (MAGIC) populations in rice: progress and potential for genetics research and breeding. *Rice (N Y)* **6** (1):11.

Basra, A. S., and C. P. Malik, 1984 Development of the Cotton Fiber, pp. 65-113.

Beasley, J. O., 1940 The Origin of American Tetraploid Gossypium Species. *American Naturalist* **74** (752):285-286.

Boquet, D. J., E. B. Moser, and G. A. Breitenbeck, 1993 Nitrogen Effects on Boll Production of Field-Grown Cotton. *Agronomy Journal* **85** (1).

Botstein, D., R. L. White, M. Skolnick, and R. W. Davis, 1980 Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms. *American Journal of Human Genetics* **32** (3):314-331.

Brush, S., R. Kesseli, R. Ortega, P. Cisneros, K. Zimmerer *et al.*, 1995 Potato Diversity in the Andean Center of Crop Domestication. *Conservation Biology* **9** (5):1189-1198.

CAES, 1960 Notice of the naming and release of a noncommercial breeding stock of cotton, C 6-5, pp. 2, edited by C.A.E. Station.

Cai, C., G. Zhu, T. Zhang, and W. Guo, 2017 High-density 80 K SNP array is a powerful tool for genotyping *G. hirsutum* accessions and genome analysis. *BMC Genomics* **18** (1):654.

Calhoun, D. S., D. T. Bowman, and O. L. May, 1997 Pedigrees of Upland and Pima Cotton Cultivars Released Between 1970 and 1995, edited by M.A.F.E. Station.

Campbell, B. T., and M. A. Jones, 2005 Assessment of genotype × environment interactions for yield and fiber quality in cotton performance trials. *Euphytica* **144** (1-2):69-78.

Campbell, B. T., V. E. Williams, and W. Park, 2009 Using molecular markers and field performance data to characterize the Pee Dee cotton germplasm resources. *Euphytica* **169** (3):285-301.

Campbell, B. T., P. W. Chee, E. Lubbers, D. T. Bowman, W. R. Meredith *et al.*, 2011 Genetic Improvement of the Pee Dee Cotton Germplasm Collection following Seventy Years of Plant Breeding. *Crop Science* **51** (3):955-968.

Campbell, B. T., P. W. Chee, E. Lubbers, D. T. Bowman, W. R. Meredith *et al.*, 2012 Dissecting Genotype × Environment Interactions and Trait Correlations Present in the Pee Dee Cotton Germplasm Collection following Seventy Years of Plant Breeding. *Crop Science* **52** (2):690-699.

Campbell, B. T., and G. O. Myers, 2015 Quantitative Genetics in *Cotton*.

Cannon, W. A., 1903 Studies in Plant Hybrids: The Spermatogenesis of Hybrid Cotton. *Bulletin of the Torrey Botanical Club* **30** (3).

Carvalho, L. P. d., F. J. C. Farias, M. M. d. A. Lima, and J. I. d. S. Rodrigues, 2014 Inheritance of different fiber colors in cotton (*Gossypium barbadense* L.). *Crop Breeding and Applied Biotechnology* **14** (4):256-260.

Chaplin, J. E., 1991 Creating a Cotton South in Georgia and South Carolina, 1760-1815. *The Journal of Southern History* **57** (2).

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

Chowdhury, K. A., and G. M. Buth, 1971 Cotton seeds from the Neolithic in Egyptian Nubia and the origin of Old World Cotton. *Biological Journal of the Linnean Society* **3** (4):303-312.

Coclanis, P. A., 1999 David R. Coker, Pedigreed Seeds, and the Limits of Agribusiness in Early-Twentieth Century South Carolina. *Business and Economic History* **28** (1):105-114.

Coclanis, P. A., 2001 Seeds of Reform: David R. Coker, Premium Cotton, and the Campaign to Modernize the Rural South. *The South Carolina Historical Magazine* **102** (3):202-218.

Cotton Incorporated, 2017 The Classification of Cotton, pp. 1-24, Cary, NC.

Crow, J. F., 1948 Alternative Hypotheses of Hybrid Vigor. *Genetics* **33**:477-487.

Culp, T. W., and D. C. Harrell, 1973 Breeding Methods for Improving Yield and Fiber Quality of Upland Cotton (*Gossypium hirsutum* L.). *Crop Science* **13** (6):686-689.

Culp, T. W., D. C. Harrell, and T. Kerr, 1979a Some Genetic Implications in the Transfer of High Fiber Strength Genes to Upland Cotton. *Crop Science* **19** (4):481-484.

Culp, T. W., A. R. Hopkins, and H. M. Taft, 1979b Breeding Insect-Resistant Cottons in South Carolina. *Technical Bulletin: S.C. Agricultural Experiment Station* **1074**:1-10.

Culp, T. W., C. C. Green, and B. U. Kittrell, 1990 Registration of Twelve Noncommercial Germplasm Lines of Upland Cotton with Resistance to Bollworm, Tobacco Budworm, and Boll Weevil. *Crop Science* **30** (1).

Dell'Acqua, M., D. M. Gatti, G. Pea, F. Cattonaro, F. Coppens *et al.*, 2015 Genetic properties of the MAGIC maize population: a new platform for high definition QTL mapping in *Zea mays*. *Genome Biology* **16**:167.

Denham, H. J., 1924 The Cytology of the Cotton Plant. I. Microspore Formation in Sea Island Cotton. *Annals of Botany* **38** (151):407-432.

Des Marais, D. L., K. M. Hernandez, and T. E. Juenger, 2013 Genotype-by-Environment Interaction and Plasticity: Exploring Genomic Responses of Plants to the Abiotic Environment. *Annual Review of Ecology, Evolution, and Systematics* **44** (1):5-29.

Dia, M., T. C. Wehner, G. W. Elmstrom, A. Gabert, J. E. Motes *et al.*, 2018 Genotype X Environment Interaction for Yield of Pickling Cucumber in 24 U.S. Environments. *Open Agriculture* **3** (1):1-16.

Fang, D. D., J. Xiao, P. C. Canci, and R. G. Cantrell, 2010 A new SNP haplotype associated with blue disease resistance gene in cotton (*Gossypium hirsutum* L.). *Theoretical and Applied Genetics* **120** (5):943-953.

Fang, H., H. Zhou, S. Sanogo, A. E. Lipka, D. D. Fang *et al.*, 2014 Quantitative trait locus analysis of Verticillium wilt resistance in an introgressed recombinant inbred population of Upland cotton. *Molecular Breeding* **33** (3):709-720.

Fang, L., X. Guan, and T. Zhang, 2017 Asymmetric evolution and domestication in allotetraploid cotton (*Gossypium hirsutum* L.). *The Crop Journal* **5** (2):159-165.

Ferguson-Smith, M. A., 2015 History and evolution of cytogenetics. *Molecular Cytogenetics* **8**:19.

Fryxell, P. A., 1963 Morphology of the Base of Seed Hairs of Gossypium I. Gross Mophology. *Botanical Gazette* **124** (3):169-199.

Gillham, F. E. M., T. M. Bell, T. Arin, G. Matthews, C. L. Rumeur *et al.*, 1995 Cotton Production Prospects for the Next Decade (Technical Paper Number 267). *The World Bank*:5-5.

Gou, J. Y., L. J. Wang, S. P. Chen, W. L. Hu, and X. Y. Chen, 2007 Gene expression and metabolite profiles of cotton fiber during cell elongation and secondary cell wall synthesis. *Cell Research* **17** (5):422-434.

Green, J. M., 1955 Frego Bract, a Genetic Marker in Upland Cotton. *Journal of Heredity* **46** (5):232-232.

Grover, C. E., H. Kim, R. A. Wing, A. H. Paterson, and J. F. Wendel, 2007 Microcolinearity and genome evolution in the *AdhA* region of diploid and polyploid cotton (*Gossypium*). *Plant Journal* **50** (6):995-1006.

Gunderson, K. L., F. J. Steemers, G. Lee, L. G. Mendoza, and M. S. Chee, 2005 A genome-wide scalable SNP genotyping assay using microarray technology. *Nature Genetics* **37** (5):549-554.

Halperin, E., and D. A. Stephan, 2009 SNP imputation in association studies. *Nature Biotechnology* **27** (4):349-351.

Hamblin, M. T., E. S. Buckler, and J. L. Jannink, 2011 Population genetics of genomics-based crop improvement methods. *Trends in Genetics* **27** (3):98-106.

Harland, S. C., 1937 Homologous Loci in Wild and Cultivated American Cottons. *Nature* **140**:467-468.

Harrell, D. C., 1974 ARS-S-30: Breeding Quality Cotton and the Pee Dee Experiment Station Florence S.C., edited by USDA.

Harris, B., 1919 Year Book and Sixteenth Annual Report. Commissioner of Agriculture, Commerce and Industries of the State of South Carolina.

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

Hosking, L., S. Lumsden, K. Lewis, A. Yeo, L. McCarthy *et al.*, 2004 Detection of genotyping errors by Hardy-Weinberg equilibrium testing. *European Journal of Human Genetics* **12** (5):395-399.

Huang, B. E., A. W. George, K. L. Forrest, A. Kilian, M. J. Hayden *et al.*, 2012 A multiparent advanced generation inter-cross population for genetic analysis in wheat. *Plant Biotechnology Journal* **10** (7):826-839.

Huang, B. E., K. L. Verbyla, A. P. Verbyla, C. Raghavan, V. K. Singh *et al.*, 2015 MAGIC populations in crops: current status and future prospects. *Theoretical and Applied Genetics* **128** (6):999-1017.

Huang, C., X. Nie, C. Shen, C. You, W. Li *et al.*, 2017 Population structure and genetic basis of the agronomic traits of upland cotton in China revealed by a genome-wide association study using high-density SNPs. *Plant Biotechnology Journal* **15** (11):1374-1386.

Hulse-Kemp, A. M., J. Lemm, J. Plieske, H. Ashrafi, R. Buyyarapu *et al.*, 2015 Development of a 63K SNP Array for Cotton and High-Density Mapping of

Intraspecific and Interspecific Populations of *Gossypium* spp. *G3 (Bethesda)* **5** (6):1187-1209.

Ingvarsson, P. K., and N. R. Street, 2011 Association genetics of complex traits in plants. *New Phytologist* **189** (4):909-922.

Islam, M. S., G. N. Thyssen, J. N. Jenkins, L. Zeng, C. D. Delhom *et al.*, 2016 A MAGIC population-based genome-wide association study reveals functional association of *GhRBB1_A07* gene with superior fiber quality in cotton. *BMC Genomics* **17** (1):903.

Jackson, E. B., and P. A. Tilt, 1968 Effects of Irrigation Intensity and Nitrogen Level on the Performance of Eight Varieties of Upland Cotton, Gossypium hirsutum L.1. *Agronomy Journal* **60** (1):13-17.

Jenkins, J. N., and W. L. Parrott, 1971 Effectiveness of Frego Bract as a Boll Weevil Resistance Character in Cotton. *Crop Science* **11** (5).

Jenkins, J. N., J. C. McCarty, O. A. Gutierrez, R. W. Hayes, D. T. Bowman *et al.*, 2008 Registration of RMUP-C5, a Random Mated Population of Upland Cotton Germplasm. *Journal of Plant Registrations* **2** (3):239-242.

Johnson, J., S. MacDonald, L. Meyer, and L. Stone, 2018 The World and United States Cotton Outlook, pp. 1-18, edited by U.S.D.o. Agriculture.

Jombart, T., S. Devillard, and F. Balloux, 2010 Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics* **11**:94.

Kohel, R. J., C. F. Lewis, J. E. Endrizzi, and E. L. Turcotte, 1984 Qualitative Genetics, Cytology, and Cytogenetics in *Cotton*, edited by E.L. Turcotte.

Kohel, R. J., 1985 Genetic Analysis of Fiber Color Variants in Cotton. *Crop Science* **25** (5):793-797.

Korte, A., B. J. Vilhjalmsson, V. Segura, A. Platt, Q. Long *et al.*, 2012 A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature Genetics* **44** (9):1066-1071.

Kovacik, C. F., and R. E. Mason, 1985 Changes in the South Carolina Sea Island Cotton Industry. *Southeastern Geographer* **25** (2):77-104.

Li, D. G., Z. X. Li, J. S. Hu, Z. X. Lin, and X. F. Li, 2016 Polymorphism analysis of multi-parent advanced generation inter-cross (MAGIC) populations of upland cotton developed in China. *Genetics and Molecular Research* **15** (4).

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

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

Li, T., X. Ma, N. Li, L. Zhou, Z. Liu *et al.*, 2017 Genome-wide association study discovered candidate genes of Verticillium wilt resistance in upland cotton (*Gossypium hirsutum* L.). *Plant Biotechnology Journal* **15** (12):1520-1532.

Liu, X., B. Zhao, H. J. Zheng, Y. Hu, G. Lu *et al.*, 2015 Gossypium barbadense genome sequence provides insight into the evolution of extra-long staple fiber and specialized metabolites. *Scientific Reports* **5**:14139.

Malosetti, M., J. M. Ribaut, M. Vargas, J. Crossa, and F. A. van Eeuwijk, 2008 A multi-trait multi-environment QTL mixed model with an application to drought and nitrogen stress trials in maize (*Zea mays* L.). *Euphytica* **161** (1-2):241-257.

Mauney, J. R., 1966 Floral Initiation of Upland Cotton *Gossypium hirsutum* L. in Response to Temperatures. *Journal of Experimental Botany* **17** (3):452-459.

Mauricio, R., 2001 Mapping Quantitative Trait Loci in Plants: Uses and Caveats for Evolutionary Biology. *Nature Reviews: Genetics* **2**:370-381.

Meredith, W. R., and R. R. Bridge, 1971 Breakup of Linkage Blocks in Cotton, *Gossypium hirsutum* L. *Crop Science* **11** (5):695-698.

Meyer, L. A., 2020 Cotton and Wool Outlook. USDA - ERS.

Minton, E. B., and R. H. Garber, 1983 Controlling the Seedling Disease Complex of Cotton. *Plant Disease* **67** (1):115-118.

Montesinos-Lopez, O. A., A. Montesinos-Lopez, J. Crossa, F. H. Toledo, O. Perez-Hernandez *et al.*, 2016 A Genomic Bayesian Multi-trait and Multi-environment Model. *G3 (Bethesda)* **6** (9):2725-2744.

Montesinos-Lopez, O. A., A. Montesinos-Lopez, J. Crossa, D. Gianola, C. M. Hernandez-Suarez *et al.*, 2018 Multi-trait, Multi-environment Deep Learning Modeling for Genomic-Enabled Prediction of Plant Traits. *G3 (Bethesda)* **8** (12):3829-3840.

Moore, J. H., 1956 Cotton Breeding in the Old South. *Agricultural History* **30** (3):95-104.

Naoumkina, M., G. N. Thyssen, D. D. Fang, J. N. Jenkins, J. C. McCarty *et al.*, 2019 Genetic and transcriptomic dissection of the fiber length trait from a cotton (*Gossypium hirsutum* L.) MAGIC population. *BMC Genomics* **20** (1):112.

Nazareno, A. G., J. B. Bemmels, C. W. Dick, and L. G. Lohmann, 2017 Minimum sample sizes for population genomics: an empirical study from an Amazonian plant species. *Molecular Ecology Resources* **17** (6):1136-1147.

NCCA, 2011 Overview of the U.S. Cotton Industry, pp. 1-45. National Cotton Council of America.

Nei, M., 1973 Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America* **70** (12):3321-3323.

Ozaki, K., Y. Ohnishi, A. Iida, A. Sekine, R. Yamada *et al.*, 2002 Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nature Genetics* **32** (4):650-654.

Palmer, S. A., A. J. Clapham, P. Rose, F. O. Freitas, B. D. Owen *et al.*, 2012 Archaeogenomic evidence of punctuated genome evolution in *Gossypium*. *Molecular Biology and Evolution* **29** (8):2031-2038.

Paterson, A. H., Y. Saranga, M. Menz, C. X. Jiang, and R. J. Wright, 2003 QTL analysis of genotype x environment interactions affecting cotton fiber quality. *Theoretical and Applied Genetics* **106** (3):384-396.

Pont, C., S. Wagner, A. Kremer, L. Orlando, C. Plomion *et al.*, 2019 Paleogenomics: reconstruction of plant evolutionary trajectories from modern and ancient DNA. *Genome Biology* **20** (1):29.

Pritchard, J. K., M. Stephens, and P. Donnelly, 2000 Inference of Population Structure Using Multilocus Genotype Data. *Genetics* **155** (2):945-959.

Raj, A., M. Stephens, and J. K. Pritchard, 2014 fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* **197** (2):573-589.

Reddy, K. R., H. F. Hodges, and J. M. McKinion, 1993 A Temperature Model for Cotton Phenology. *Biotronics* **22**:47-52.

Reinisch, A. J., J. Dong, C. L. Brubaker, D. M. Stelly, J. F. Wendel *et al.*, 1994 A Detailed RFLP Map of Cotton, *Gossypium hirsutum x Gossypium barbadense*: Chromosome Organization and Evolution in a Disomic Polyploid Genome. *Genetics* **138**:839-847.

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

Rungis, D., D. Llewellyn, E. S. Dennis, and B. R. Lyon, 2005 Simple sequence repeat (SSR) markers reveal low levels of polymorphism between cotton (*Gossypium hirsutum* L.) cultivars. *Australian Journal of Agricultural Research* **56** (3).

Shim, J., P. K. Mangat, and R. B. Angeles-Shim, 2018 Natural Variation in Wild *Gossypium* Species as a Tool to Broaden the Genetic Base of Cultivated Cotton. *Journal of Plant Science Current Research* **2** (005):1-9.

Shimshi, D., and A. Marani, 1971 Effects of Soil Moisture Stress on Two Varieties of Upland Cotton in Israel II. The Northern Negev Region. *Experimental Agriculture* **7** (3):225-239.

Skovsted, A., 1933 Cytological Studies in Cotton. I. The Mitosis and the Meiosis in Diploid and Triploid Asiatic Cottons. *Annals of Botany* **47** (186):2270251.

Skovsted, A., 1934 Cytological Studies in Cotton II. Two Interspecific Hybrids Between Asiatic and New World Cottons. *Journal of Genetics* **28**:4077-4424.

Skovsted, A., 1937 Cytological Studies in Cotton IV. Chromosome Conjugation in Interspecific Hybrids. *Journal of Genetics* **34** (1):97-134.

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

Smith, A. L., 1964 Registration of Auburn 56 Cotton (Reg. No. 45). *Crop Science* **4** (4).

Smith, C. W., and J. T. Cothren, 1999 *Cotton: Origin, History, Technology, and Production*: John Wiley & Sons.

Stephens, S. G., 1955 Linkage in Upland Cotton. *Genetics* **40** (6):903-917.

Stephens, S. G., 1976 Some observations on photoperiodism and the development of annual forms of domesticated cottons. *Economic Botany* **30** (4):409-418.

Stewart, J. M. D., 1975 Fiber Initiation on the Cotton Ovule (*Gossypium hirsutum*). *American Journal of Botany* **62** (7):723-730.

Sun, Z., X. Wang, Z. Liu, Q. Gu, Y. Zhang *et al.*, 2017 Genome-wide association study discovered genetic variation and candidate genes of fibre quality traits in *Gossypium hirsutum* L. *Plant Biotechnology Journal* **15** (8):982-996.

Tabery, J., 2008 R. A. Fisher, Lancelot Hogben, and the origin(s) of genotype-environment interaction. *Journal of the History of Biology* **41** (4):717-761.

Thomas, D. H., 1965 Pre-Whitney Cotton Gins in French Louisiana. *The Journal of Southern History* **31** (2).

Thyssen, G. N., D. D. Fang, R. B. Turley, C. Florane, P. Li *et al.*, 2014 Next generation genetic mapping of the Ligon-lintless-2 (*Li(2)*) locus in upland cotton (*Gossypium hirsutum* L.). *Theoretical and Applied Genetics* **127** (10):2183-2192.

Thyssen, G. N., J. N. Jenkins, J. C. McCarty, L. Zeng, B. T. Campbell *et al.*, 2019 Whole genome sequencing of a MAGIC population identified genomic loci and candidate genes for major fiber quality traits in upland cotton (*Gossypium hirsutum* L.). *Theoretical and Applied Genetics* **132** (4):989-999.

Turner, J. H., 1974 ARS-W-16: History of Acala Cotton Varieties Bred for San Joaquin Valley, California, pp. 1-23, edited by U.S.D.o.A.-A.R. Service.

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

van Eeuwijk, F. A., M. C. Bink, K. Chenu, and S. C. Chapman, 2010 Detection and use of QTL for complex traits in multiple environments. *Current Opinion in Plant Biology* **13** (2):193-205.

VanLiere, J. M., and N. A. Rosenberg, 2008 Mathematical properties of the r2 measure of linkage disequilibrium. *Theoretical Population Biology* **74** (1):130-137.

Wang, H., K. P. Smith, E. Combs, T. Blake, R. D. Horsley *et al.*, 2012a Effect of population size and unbalanced data sets on QTL detection using genome-wide association mapping in barley breeding germplasm. *Theoretical and Applied Genetics* **124** (1):111-124.

Wang, K., Z. Wang, F. Li, W. Ye, J. Wang *et al.*, 2012b The draft genome of a diploid cotton Gossypium raimondii. *Nature Genetics* **44** (10):1098-1103.

Wang, K., J. F. Wendel, and J. Hua, 2018 Designations for individual genomes and chromosomes in *Gossypium*. *Journal of Cotton Research* **1** (1).

Wang, M., L. Tu, M. Lin, Z. Lin, P. Wang *et al.*, 2017a Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication. *Nature Genetics* **49** (4):579-587.

Wang, R., A. Stec, J. Hey, L. Lukens, and J. Doebley, 1999 The limits of selection during maize domestication. *Nature* **398**:236-239.

Wang, W. W., Z. Y. Tan, Y. Q. Xu, A. A. Zhu, Y. Li *et al.*, 2017b Chromosome structural variation of two cultivated tetraploid cottons and their ancestral diploid species based on a new high-density genetic map. *Scientific Reports* **7** (1):7640.

Ware, J. O., 1932 Inheritance of Lint Colors in Upland Cotton. *Agronomy Journal* **24**:550-562.

Ware, J. O., 1937 Plant Breeding and the Cotton Industry in *Year Book of Agriculture*. USDA.

Wells, E., 2019 South Carolina County Estimates Cotton 2017-2018, edited by N.S.D. USDA.

Wendel, J. F., P. D. Olson, and J. M. D. Stewart, 1989 Genetic Diversity, Introgression, and Independent Domestication of Old World Cultivated Cottons. *American Journal of Botany* **76** (12):1795-1806.

Wendel, J. F., 1989 New World tetraploid cottons contain Old World cytoplasm. *Proceedings of the National Academy of Sciences of the United States of America* **86**:4132-4136.

Wendel, J. F., and V. A. Albert, 1992 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* **17** (1):115-143.

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

Wright, S., 1965 The Interpretation of Population Structure by F-Statistics with Special Regard to Systems of Mating. *Evolution* **19** (3):395-420.

Wurschum, T., 2012 Mapping QTL for agronomic traits in breeding populations. *Theoretical and Applied Genetics* **125** (2):201-210.

Xiao, J., D. D. Fang, M. Bhatti, B. Hendrix, and R. Cantrell, 2010 A SNP haplotype associated with a gene resistant to *Xanthomonas axonopodis* pv. *malvacearum* in upland cotton (*Gossypium hirsutum* L.). *Molecular Breeding* **25** (4):593-602.

Yapa, L., 1993 What are Improved Seeds? An Epistemology of the Green Revolution. *Economic Geography* **69** (3):254-273.

Yu, J., J. B. Holland, M. D. McMullen, and E. S. Buckler, 2008 Genetic design and statistical power of nested association mapping in maize. *Genetics* **178** (1):539-551.

Yuan, D., Z. Tang, M. Wang, W. Gao, L. Tu *et al.*, 2015 The genome sequence of Sea-Island cotton (*Gossypium barbadense*) provides insights into the allopolyploidization and development of superior spinnable fibres. *Scientific Reports* **5**:17662.

Zhang, J., A. Abdelraheem, G. N. Thyssen, D. D. Fang, J. N. Jenkins *et al.*, 2020 Evaluation and genome-wide association study of Verticillium wilt resistance in a MAGIC population derived from intermating of eleven Upland cotton (*Gossypium hirsutum*) parents. *Euphytica* **216** (1).

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

Zhao, F., and S. Xu, 2012 Genotype by environment interaction of quantitative traits: a case study in barley. *G3 (Bethesda)* **2** (7):779-788.

CHAPTER TWO

POPULATION STRUCTURE AND GENETIC DIVERSITY OF THE PEE DEE COTTON GERMPLASM COLLECTION

Abstract

Accelerated marker-assisted selection and genomic selection breeding systems require high quality genotyping data for parental material to optimally allocate breeding resources. Since 1935, the Pee Dee cotton germplasm enhancement program has developed an important genetic resource for upland cotton (*Gossypium hirsutum* L.) contributing genetics for improved fiber quality, agronomic performance, and genetic diversity. To date, a detailed genetic survey of the program's eight historical breeding cycles has yet to be undertaken. The objectives of this study were to evaluate genetic diversity across and within breeding groups, examine population structure, and contextualize these findings relative to the global upland cotton gene pool. The CottonSNP63K array was used to identify 17,441 polymorphic markers (unthinned) in a panel of 114 diverse Pee Dee genotypes. A subset of 4,597 markers was selected to decrease marker density bias. Identity by state (IBS) pairwise distance varied substantially, ranging from 0.55 to 0.97. Pedigree-based estimates of relatedness were lowly predictive overall of the observed genetic distances. Few rare alleles were present, with 99.1% new alleles appearing within the first four breeding cycles. Population structure analysis with principal component analysis, discriminant analysis of principal components, fastSTRUCTURE, and phylogenetics revealed an admixed population with moderate substructure. Allele frequency analysis indicated potential selection signatures

associated with biotic stress resistance. The results of this study will steer future utilization of our program's germplasm resources, aid in combining program-specific beneficial alleles and maintaining genetic diversity, and establish the basis for genomic selection.

Introduction

The Pee Dee (PD) cotton germplasm enhancement program in Florence, South Carolina, was formalized in 1935 as part of the USDA Agricultural Research Service's goal to revitalize Sea Island cotton (*Gossypium barbadense* L.) cultivation (Harrell 1974). Over time, the PD program transitioned into an Upland cotton (*Gossypium hirsutum* L.) long-term breeding effort, focused on the improvement of fiber strength and other quality traits, insect resistance, and other key agronomic traits (Campbell et al. 2011). Complex intercrossing, mating schemes, and germplasm recycling have led to the development of unique breeding materials and cultivars throughout the history of the program (Culp 1998). Sources of genetic diversity for the PD program include accessions include *G. barbadense, G. hirsutum*, and the triple hybrid series composed of genes from *G. hirsutum*, *G. arboreum* L., and *G. thurberi* Tod. (Beasley 1940). Germplasm releases from PD have been distributed and utilized across many public and private cotton breeding programs, especially as a source for combined fiber length and strength (Bowman and Gutierrez 2003; Calhoun et al. 1997)

From 1935 to 2000, the PD program completed eight breeding cycles, generating groups of cultivars and germplasm lines in each cycle (Campbell et al. 2011). Group one started with the crossing of founding parents to generate new intercrossed, recombinant lines with interspecific (combination of genetic backgrounds from multiple species) sources of fiber length and strength alleles. Groups two, three, and four were developed through the intercrossing of materials generated in the first three cycles. Groups five and six represented a change in breeding objectives as efforts were made to develop host plant resistance to the boll weevil (*Anthonomus grandis* Boh.). Group seven began another change in the PD program, where materials from outside of the breeding program were incorporated as breeding parents in an effort to bring new sources of genetic variation for increased yield potential. Group eight resulted from a combination of intercrossing of materials developed in prior breeding cycles, along with the introduction of more breeding parents from outside the PD program. The program's history is summarized graphically in **Figure 2.1**.

A retrospective accounting of the breeding resources produced from the program over its 85-year history was undertaken to better understand the breeding history of the PD program and to aid us in efforts to accelerate present breeding efforts. In 2009, data from a multi-site-year field experiment was combined with 80 polymorphic simple sequence repeat (SSR) markers to characterize the phenotypic and genetic variability across these eight breeding groups (Campbell et al. 2009). They found variability for multiple fiber quality and yield components, including fiber length, fiber strength, fiber fineness, and lint percent, among others. However, the study was limited by molecular markers and genotyping techniques available at the time. Modern genotyping technologies, like the CottonSNP63K array released in 2015 (Hulse-Kemp et al. 2015), have enabled a host of new experiments and discoveries in cotton.

Population structure and diversity, assessed by the scoring of genome-wide genetic markers such as single nucleotide polymorphisms (SNPs), is crucial to generating an unbiased picture of the genomic landscape before undertaking genome-wide association studies (GWAS) or genomic selection (Hamblin et al. 2011). Multiple methods are available for evaluating population structure, ranging from the classic phylogenetic model, which uses hierarchical clustering on the genetic distance matrix to identify similar and different members of a population (Odong et al. 2011). Principal component analysis has long been used to correct for population structure in further genomic analyses (Price et al. 2006). Other methods, such as discriminant analysis of principal components and fastSTRUCTURE, enable the visualization and evaluation of complex stratification in such panels as nested association mapping or breeding populations (Jombart et al. 2010; Raj et al. 2014; Huang et al. 2015; Maurer et al. 2015; Deperi et al. 2018).

Marker-trait association experiments have resulted in the discovery of dozens of quantitative trait loci (QTL) underlying diverse traits including salt tolerance, fiber quality, and wilt resistance (Sun et al. 2018; Gapare et al. 2017; Abdelraheem et al. 2020). Efforts to characterize the genetic diversity and population structure in the US upland cotton gene pool have also been undertaken. Tyagi et al. (2014) used a set of 122 polymorphic SSR marker, which were able to successfully distinguish 378 cultivars and breeding lines originating from the western, southwestern, midsouth, and eastern US cotton growing regions. They observed similar correspondence between PCA, STRUCTURE, and allele frequency methods, noting an overall low level of genetic diversity relative to other crop species. Hinze et al. (2017) evaluated germplasm from the upland cotton core collection, with a focus on comparing a catalogue of phenotypic traits to SNP genotypes from the CottonSNP63K array. Multidimensional scaling analysis revealed overlap between germplasm originating from the US and other places in the world, and a moderate ability to distinguish germplasm by US cotton growing region. However, they did not observe meaningful clustering within improved upland cotton germplasm with the fastSTRUCTURE method.

The goal of this study was to evaluate genetic diversity across and within PD breeding groups and relate these findings to the worldwide improved upland cotton germplasm. We hypothesized that this closed (largely inbreeding) breeding program, with long breeding cycles, complex intermating, and multiple shuffling of potentially unique alleles would provide an interesting population genetics model for studying the effects of genetic drift and artificial selection. Hence, the objectives of this study were to evaluate genetic diversity across and within PD program breeding groups by utilizing genome-wide SNP markers from the Cotton SNP63K array, examine population structure, and contextualize these findings relative to the global upland cotton gene pool.

Materials and Methods

Description of Plant Genotypes and Genotyping

Representative plant genotypes from each of eight PD breeding groups were selected for examination, covering 96 released breeding lines and cultivars (**Figure 2.1**). Seeds were requested from the US National Cotton Germplasm Collection in College Station, TX, and grown in a greenhouse in Florence, SC, during Winter 2018. Three seeds for each genotype were planted and thinned to a single plant at the cotyledon stage. Newly emerged leaves were collected in 1.5 ml centrifuge tubes and immediately placed on ice. Leaf tissue was stored at -80C until processing for DNA extraction. Frozen leaves were lysed in a tissue homogenizer with two added glass beads. Genomic DNA extraction was performed using the DNeasy Plant Mini Kit (Qiagen Inc, Germantown, MD, USA) according to manufacturer instructions. Sample DNA concentration was measured using a NanoDrop Spectrophotometer (Thermo Fisher Scientific Inc, Waltham, MA, USA). A vacuum centrifuge was used to concentrate samples with concentration below 100 ng/µl. Samples of 25 µl were loaded onto a 96-well plate and shipped on dry ice overnight to the Texas A&M Institute for Genomic Sciences and Society (College Station, TX, USA). Upon receipt, samples were quality checked at Texas A&M using the PicoGreen assay (Ahn et al. 1996), and adjusted to a DNA concentration of 50 ng μl^{-1}. Standardized DNA samples were hybridized with the CottonSNP63K array, a custom Infinium iSelect HD Genotyping Assay (Illumina Inc., San Diego, CA), developed by Hulse-Kemp et al. (2015). The standard cluster file and output parameters were employed for export to a plain text final report file (https://www.cottongen.org/data/community_projects/tamu63k#T1). The final report file from Illumina GenomeStudio was filtered using a custom Python script, retaining only markers listed as Functional Polymorphic (Hulse-Kemp et al. 2015), by minor allele frequency (MAF > 2.5%) and call rate (CR > 90%) to generate Dataset One. Marker probe sequences were mapped to the UTX_v2.1 reference genome (Chen et al. 2020).

The filtered data was ported to the plink data format for compatibility with plink 1.9 (Chang et al. 2015). A SNP matrix in the -1/0/1 format was also generated for use with some R packages. Putative linkage disequilibrium blocks were discovered with the "--indep-pairwise" command in plink 1.9 (Chang et al. 2015).

The SNP data of 267 improved upland cotton (*Gossypium hirsutum*) samples genotyped on the CottonSNP63K Array were downloaded from the array project page on CottonGen (Yu et al. 2014) and converted to PED format using a custom python script. Duplicated genotypes (IBS > 97%) were excluded from further analysis. A total of 249 improved upland cotton lines (non-PD lines) were included in the analysis, as well as 114 PD lines (96 from the present study and 18 from CottonGen). Markers were filtered to include those with MAF > 2.5% and CR > 90%. Summary statistics were calculated for the number of markers passing filtering during each step using the "--het" and "--freq" commands in plink. Percent heterozygosity for each individual in each dataset was also calculated by dividing the number of observed heterozygous calls by the total number of calls.

Population Structure Analysis

Breeding group designations were selected based on parentage and the breeding history of the PD program (Campbell et al. 2011). These group designations were used as the prior (assumed) group designation for naïve population structure analyses. Two principal component analysis (PCA) variants were tested to identify a consensus idea of clustering with and without thinning and between classic PCA and double-centered PCA. The first variant used was classic PCA in plink, which itself is a direct port of the PCA function in EIGENSTRAT (Price et al. 2006), with normal reference/alternate allele coding and a built-in method to handle missing data. The second variant used was double-centered PCA as implemented by a custom R script which did not include a mechanism for missing values; therefore, the median SNP value for each locus was used to replace no-calls. The double-centered PCA run included an additional pre-processing step, which included changing minor allele coding to homozygous major allele=0, heterozygous=1, and homozygous minor allele=2 (Gauch et al. 2019). Biplots of individuals for the SNP x Individual interaction were generated for datasets one and two with individuals color coded by the prior breeding group number. Eigenvalues were used to calculate percent variance explained by the first two dimension of PC, calculated as the eigenvalue for the eigenvector divided by the sum of the eigenvalue for the first 40 eigenvectors. To reduce the effect of sign changing on the visual interpretation of PCA biplots, the PC1 vector was flipped by multiplying by -1 when necessary (Gauch et al. 2019).

To test for differences between breeding groups, discriminant analysis of principal components (DAPC) was performed on Dataset Two with the R package adegenet (Jombart et al. 2010). Prior group assignment was based on the original breeding cycle assignment. The plink bed format file was converted to a genind object using the "genomic_converter" function in the R package "radiator" (Gosselin et al. 2020). A plink raw file, generated with the "--recode A" flag, was read in together with the map file with the "read.plink" command as a genlight object. DAPC was performed naively with the "dapc" command in interactive mode. To avoid model overfitting, the "optim.a.score" function was used to select the number of principal components. A DAPC biplot was generated using the original group numbers. The "compoplot.dapc" function was used to calculate and graph the assignment of individuals to each of the eight breeding groups .

Population structure was also evaluated with the maximum likelihood tree in MEGA X (Kumar et al. 2018). The concatenated DNA fasta file was generated by exporting from ped format with PGDSpider (Lischer and Excoffier 2012) and reading into MEGA X. The best DNA model was chosen using the minimum Bayesian information criterion "Find Best DNA/Protein Model" without invariable sites. A test of phylogeny was then performed with the optimal model, the general time reversible model, and the bootstrap method with 1000 replicates. Branches with less than 50% bootstrap support were collapsed into polytomies. The tree was plotted as a phylogram with the "plot.phylo" function in the R package "ape" (Paradis and Schliep 2019).

To test for the number of groups and group membership of each genotype, the "chooseK.py" function in fastSTRUCTURE was used for k=1-10 (Raj et al. 2014). The diagram for fastSTRUCTURE results was made by converting to a matrix object in R and plotting using the "compoplot" function in adegenet. To identify DAPC clusters, the "find.clusters.genlight" command was used, with 40 PCs retained. The number of DAPC-derived groups was chosen using the minimum value of the Bayesian information criterion. These identified clusters from DAPC were retained and plotted in a Sankey diagram to examine the relationship between the three classification methods.

Signatures of Selection in the PD Program

To test for putative signatures of selection in the PD program versus other improved Upland cotton genotypes, a marker-specific Bayes factor (BF), analogous to Wright's F_{ST}, was estimated for each marker with the function in BayEnv2 (Coop et al. 2010; Gunther and Coop 2013). Samples were classified as PD or World (non-PD). The log10 of resulting BFs were plotted in a manhattan plot with a threshold of $\log10(BF) > 2$. Allele frequency plots for the each of the significant markers were also generated. Putative regions under selection were determined as chromosomal segments containing significant markers (BF > 10). A list of genes and their gene ontology terms in these regions was identified using the GFF3 annotation file for the annotation of the *Ghir* reference genome assembly (Chen et al. 2020). The list of genes was subjected to gene enrichment analysis with the weight-count method ($p < 0.05$) and ranked by Fisher's exact test with the R package 'topGO' (Alexa and Rahenfuhrer 2020).

Gene Enrichment Analysis for Breeding Groups Five and Six vs All

Each of the 114 PD genotypes were assigned to one of three clusters based on DAPC. The cluster containing mostly genotypes from breeding groups five and six was assigned as one group for selection analysis, and all other genotypes were assigned to another group. Selection analysis was performed to compare between the groups five and six cultivars against all others.

Results

Dataset One, the filtered set of markers, contained 17,441 markers anchored to a position on the *Ghir* v2.0 reference genome (Chen et al. 2020). During filtering, an initial set of 38,869 known polymorphic markers across any *Gossypium* spp. had 19,952 markers excluded with MAF < 2.5%, 280 markers excluded with CR < 90%, and 1,196 were excluded due to no available reference genome position38. After thinning to account for marker redundancy due to high linkage disequilibrium (LD), Dataset Two reduced this number to 4,597 markers. The marker density across 15 of the 26 chromosomes differed significantly between Dataset One and Two (**A**). In Dataset One, the number of markers ranged from 1,629 on chr A08 to 247 on chr A02. After thinning to account for variable marker density, the number of markers per chromosome was more uniform, ranging from a maximum of 268 SNPs on chr D05 to 116 on chr A02 (**Supplemental Figure 2.1 B**). Overall, the changes corresponded to a reduction in SNP overrepresentation in low recombination pericentromeric regions.

Of the 9,194 alleles (two alleles for each of 4,597 SNPs) present in at least two of the 114 individuals in Dataset Two, 95% were introduced, or detected in at least one individual, in group one, 2.9% in group two, 1.1% in group three, 0.5% in group four, and <0.4% in each of groups five through eight, indicating that most of the genetic diversity present in the PD germplasm pool was introduced in the first few cycles of breeding development. Most SNP alleles were present in at least two groups. However, group eight contained five unique SNP alleles, two of which flank a haploblock present in the denser set of variants in Dataset One, corresponding to a run of 17 group-unique alleles in 408 kb region of chr A05 (109.48 - 109.89 Mb). Heterozygosity varied substantially between genotypes (**Supplemental Table 2.2**), meaning few SNPs were completely fixed in any breeding group . Of the 9,194 alleles, 457 alleles were fixed (present in at least one copy in every genotype) in breeding group one, 764 in group two, 854 in group three, 702 in group four, 816 in group five, 561 in group six, 569 in group seven, and 273 in group eight.

Both datasets exhibited similar distributions of identity-by-state (IBS) scores. The mean pairwise genetic distance was highly similar, 0.661 in Dataset One and 0.665 in Dataset Two. Pairwise IBS genetic distances ranged in Dataset Two from 0.553 for Sealand-3 (AHK) and Sealand-542 (AHK), the two most dissimilar individuals, to 0.967 for PD 762 and PD 948, the two most similar individuals. Comparison of the additive genetic relationship matrix derived from these two datasets, which is analogous to IBS distance except it ranges from around zero to a maximum of two, also indicated high concordance (**Supplemental Figure 2.2**). When compared to the generalized numerator relationship matrix from NumericwareN, which is the comparable estimate from pedigree data, the values calculated from Dataset Two were in higher agreement ($R^2 = 0.20$) than those of Dataset One ($R^2 = 0.13$) with the pedigree-based scores (**Supplemental Figure 2.3**). Average within group genetic distances were generally higher (*ie,* pairs were more similar) than between group comparisons (**Table 2.1**).

Both PCA and double-centered PCA both showed similar results across the two datasets with the exception of PCA on Dataset One (**Figure 2.2**). Normal PCA and double-centered PCA supported the same relationship between breeding groups in Dataset Two. To mitigate for the effect of variable marker density across the chromosomes, further analyses on the PD genotypes was performed with only Dataset Two.

Both fastSTRUCTURE and phylogenetic analysis were consistent across both datasets, so the output from Dataset Two is discussed here. The results from fastSTRUCTURE supported the existence of multiple groups (k = 5), and 55 of 114 individuals were classified at the ≥80% level of probability (**Figure 2.3**). De novo group assignments, either through DAPC or fastSTRUCTURE, supported the original eight groups with the novel groups representing a superset, or overlap, of the predicted breeding groups (**Figure 2.4**). The consensus phylogenetic tree also identified the same subgroups as fastSTRUCTURE and DAPC (**Figure 2.5**).

To explore the genetic differentiation of the PD germplasm (PD Group) from other improved *G. hirsutum* cultivars (World Group), a Bayes factor was calculated to compare genetic differentiation relative to the background level of genetic differentiation between the groups at each of 20,566 polymorphic SNPs. The Bayes factor was log10-transformed and plotted for each SNP, with allele frequencies at six interesting SNPs for the eight breeding groups and world group plotted (**Figure** 2.6). The 36 putative SNP markers under selection were located at 32 genetic locations, distributed across 13 chromosomes. These regions contained 118 genes, enriched for gene ontology (GO) terms related to response to stimuli, translation, actin, and glutathione metabolic process (**Table 2.2**).

Discussion

We hypothesized that a SNP survey of 114 representative individuals from the PD cotton germplasm enhancement program would reflect population structure over eight breeding cycles, spanning more than 85 years of breeding. The CottonSNP63K platform provided an efficient and repeatable method for identifying 17,441 high-quality, polymorphic SNP markers in the PD cotton germplasm. Due to the relatively closed nature of the breeding program, we expected that large haploblocks could complicate estimates of population structure and relatedness. To compensate for these LD patterns, a thinned dataset was generated to ensure that long haploblocks segments would not bias our analysis. The thinned dataset performed surprisingly similarly to the higher-density SNP set that included more than four times as many markers, indicating that lower density genotyping may have provided an equivalently robust basis for evaluation of population structure.

The effect of thinning on the interpretation of SNP data was first evaluated by comparing the additive genetic relationship matrix (GRM) between Dataset One and Two, which exhibited strong agreement ($R^2 = 0.77$ - **Supplemental Figure 2.2**). However, when fit to the pedigree-based relationship estimate, pairwise comparisons calculated from Dataset Two ($R^2 = 0.19$) fit the expectation better than those for Dataset One ($R^2 = 0.09$ **- Supplemental Figure 2.2**). Because thinning in Dataset Two reduced the high weight from redundant alleles, the dispersion of the GRM was higher in Dataset One (SD = 0.044) than Dataset Two (SD = 0.036). The lower dispersion of scores from Dataset Two may have contributed to better fit to the pedigree-based scores.

Pee Dee breeding groups one through four have common parentage composed of approximately 12 diverse founders (Culp et al. 1979). Indeed, most of the allelic diversity was introduced in the first four breeding groups, accounting for 99.5% of the total SNP alleles in Dataset Two present in this closed breeding program. This apparent lack in allelic diversity was compensated by the complex combinations of these alleles across the history of the program. For some genotype pairs we hypothesized a high level of genetic similarity; however, some re-selection pairs of lines, published as separate germplasm releases purportedly from the same gene pool, were more genetically distinct than other completely unrelated pairs. For example, 'PD-3' and PD-3-14, released as a reselection of PD-3, had a pedigree-based kinship ~1.00 but a genetic distance of 0.76, indicating they were only somewhat more different from each other than the average pair of genotypes. Regardless, the average IBS genetic distance of genotype pairs, ~0.66, similar to ~0.71 for improved upland cotton according to Hinze et al. (2017), and ~0.80 for Tyagi et al. (2014). The variable estimates reflect differing numbers of genetic markers types, population sizes, distribution of markers, type of plant genotypes used in the study (*ie*, obsolete vs elite), and differences in how rare alleles change genetic distance.

We hypothesized that within-breeding group genetic variation would be lower than between-breeding group variation, since members of a breeding group tended to have similar selection regimes and parents (**Table 2.1**). Given the IBS distance scores calculated from Dataset Two, with the exception of breeding groups one and five, individuals within the group were on average more similar to one another than with members of another group, which supported our hypothesis. Interestingly, individuals within groups two and three were more similar to breeding group one than they were to each other, perhaps indicating additional selection and/or drift among genotypes in these groups. This was reflected in the DAPC where *k*=3 and genotypes from breeding groups one, two, and three were primarily placed within the same cluster (**Figure 2.4**).

Pairwise genetic distance alone was inadequate to fully capture the genetic diversity present within and between breeding groups. Both methods of PCA analysis (PCA and double-centered PCA) for Datasets One and Two had surprisingly consistent results, considering the number of SNPs changed by a factor four. Principal component analysis, when applied to genome-wide data, is able to capture underlying genetic structure by summarizing the differences between individuals at the SNP by individual interaction level (Price et al. 2006; Gauch et al. 2019). In all cases, once flipped for sign changes in PC1, the primary dimension of PC showed a gradient of separation between the earlier groups, one through four, on one extreme (**Figure 2.2**). The host-plant insect resistant breeding groups, five and six, were in the other extreme; and the most recent groups, seven and eight, were in the middle. The primary dimension, PC1, explained between 10.6% and 13.1% of the variance included in the first 40 PCs. The second dimension, PC2, was the same for all plots except for the plink PCA of the unthinned Dataset One. In all other plots, the newer groups, seven and eight, clustered together on one pole and the other six groups in the other pole. In the plink PCA biplots, the group separation was not apparent in PC2, with five outlier individuals present at one extreme and all other individuals clustering together at the bottom.

The outliers for the unthinned plink PCA plots in PC2 included PD 3246 (AC 239/FJA 348), PD 9232 (‘Coker 421’/ PD 2164), PD 93034 (PD 5285/PD 5485), PD 93004 (Brown Accession/PD -3) and Sealand 3 (resel. ‘Sealand’: ‘Coker Wilds’/’Bleak

Hall') at the furthest extreme, and PD 93001 (Brown Accession/PD-3) and PD 5576 ('Deltapine 41'/PD 3246) near the center of the two large clusters. Two of these individuals are brown lint cottons, PD 93001 and PD 93004. PD 3246 is the pollen donor for the original cross for PD 5576 and is also a full sib of PD 2164, one of the parents of PD 9232. Although these lines were outliers in this analysis, there were other individuals in the study with highly similar parentage and selection strategies, suggesting that common pedigrees and brown lint do not alone explain these outliers.

The loadings for variant weights in PC2 of plink PCA for Dataset One revealed significant contribution (27.8% of total variant loadings) from a run of 911 markers in high LD on chromosome A08 (16.46 Mb to 79.48 Mb). After thinning based on putative haploblocks, this segment was reduced to include only 21 markers. Pedigree analysis indicated a possible common breeding program origin for this chromosomal segment from 'Hopi Moencopi' via C-6-5, a California breeding line used early in the development of the PD program. Another potential origin is Coker Wilds or Bleak Hall via Sealand. Interestingly, the pericentromeric region of chr A08 has been noted as exhibiting low recombination frequency (Shen et al. 2017), which may be due to gametic incompatibility associated with multiple large scale inversions in this region of chr A08 (Yang et al. 2019). The two individuals near the center of the two major clusters in PC2, PD 93001 and PD 5576, were heterozygous for >90% of these 911 markers, indicating a potential region of fixed heterozygosity. These regions accounted for a 70% and 27% increase in observed heterozygosity for PD 93001 and PD 5576, respectively, between Dataset One and Two (**Supplemental Table 2.2**). Five other individuals from the improved germplasm set (‘Coker 315’, ‘Reba P279’, ‘Acala 5’, ‘Lockett BXL’, and ‘Deltapine 16’) shared this region of heterozygosity. All other individuals were >95% homozygous in this region, except for ‘Sicala-3-2’ and ‘Namcala’ which had a high number of no-calls in this region. Fifty-one of the 249 improved Upland cotton samples from CottonGen are homozygous for the minor haplotype.

The other three PCA biplots, however, show a much clearer picture of the interrelatedness of individuals. An arc of individuals is present, spanning from those with low values in PC1 and PC2, near zero values for PC1 and high PC2, and those with high values in PC1 and low values in PC2. The central cluster was mostly composed of individuals from groups seven and eight, with overlap on the left of groups one through four and on the right groups five and six. Examination of variant weights did not indicate highly weighted genomic regions, a potential indicator of bias as the case had been with plink PCA, suggesting that polymorphism across the genome was responsible for separation between individuals. Plots of additional dimensions of PCA did not reveal any obvious structure relative to the original breeding group classifications (data not shown).

One possible biological interpretation of these results is that PC1 and PC2 captured two allele frequency gradients (Novembre and Stephens 2008). The primary axis, PC1, may involve alleles associated with high frequency in breeding groups five and six, perhaps associated with the genetic background of their parents. In this model, the earlier breeding groups may have low levels of this genetic background, the newest groups seven and eight have moderate levels, and groups five and six have the highest frequency. Indeed, gene enrichment analysis revealed genes nearby genetic markers associated with the separation between breeding groups five and six and other PD genotypes associated with the citric acid cycle, aerobic respiration, and steroid biosynthesis and metabolism. Two genes with one of the putative chromosomal segments under selection (chr A03 97.260 Mb to 97.282 Mb), *A05G350300* and *A05G350400*, are tandem-repeat homologs of an *Arabidopsis thaliana* gene annotated as 2-oxoglutarate dehydrogenase, E1 component. The complex of this gene product has been implicated in plant immunity response via salicylic acid, suggesting a potential role in host plant resistance (Klessig et al. 2016). The secondary axis, PC2, may involve the frequency of SNP alleles associated with elite, modern cultivars, with individuals from groups seven and eight having the highest frequency of these alleles.

Another possibility is that the plink PCA plots of the unthinned Dataset One reveals the "true" population structure and the other three plots are examples of PCA "arch distortion." Arch distortion results from the projection of a single gradient onto the first two, dominant dimensions of PCA (Gauch et al. 2019). For example, perhaps PC1 and PC2 in the other six PCA plots are simply capturing the same information as PC1 in the other two plots. However, these six plots do not have the characteristic closed arch at the bottom of the plot, and both dimensions have plausible biological interpretations.

Additional support for this two-gradient hypothesis is found in the results from DAPC. These results project a summary of principal components onto the groups rather than individuals, thus minimizing error relative to the group assignments rather than individuals (Jombart et al. 2010). Hence, DAPC explores differences within and between groups while traditional PCA is optimized for differences across all individuals. The

DAPC biplot, when tuned to the number of PCs included to reduced model overfitting, shows the same relationship between breeding groups, and individuals within those groups, similar to that in plink PCA and double-centered PCA (**Supplemental Figure 2.4**). Individuals in each group cluster close to each other, with groups two having the widely spread individuals, which is consistent with the group two having the highest within-group genetic diversity based on average pairwise IBS genetic distance. However, the trend for the average position of each group is much more obvious. From group one through four, as the average individual is progressively more "improved" in terms of breeding and quality, they plot closer to groups seven and eight. Additionally, the less improved pest-resistant group five, as compared to group six, has individuals more widely dispersed in the vertical axis, whereas group six is even further from the other groups as selection pressure for insect resistance scaled up across generations.

While the breeding group classification system provides a historic starting point for understanding the structure of breeding program material, it cannot alone account for the effects of genetic drift, selection, and/or outcrossing. In the hope of revealing lasting signatures of these dynamics, we plotted populations and membership probabilities for each genotype identified using fastSTRUCTURE (**Figure 2.3**) given the highest likelihood number of subgroups (*k*=6) before the fit value oscillated nearby. The results from fastSTRUCTURE were consistent with our expectations given the breeding history for surveyed genotypes. Population One include three of the 'Sealand' germplasm lines, resulting from the cross between Coker Wilds and Bleak Hall. Population Two is composed entirely of founding lines and intercrosses between them. Population Three includes mostly early crosses between founding lines and elite introductions ‘Coker 421’, ‘MO-DEL’, and ‘AU-56’. Population Four includes PD 695, PD 875, and 18 selections from their progeny, all sharing a common grandparent LA Frego 2, an insect-resistant frego-bract line. We identified PD 2165 (PI 529618) as an outlier in Population Four, whereas the other version of PD 2165 (PI 529242) included in this analysis clustered with earlier releases as expected. This confirms that the two PD 2165 entries obtained from the Germplasm Resource Information Network (USDA-ARS 2015) and used in this study represent different genotypes. A definitive explanation for this difference is not known; however, it is likely that one of the two versions (likely PI 529618) was mislabeled or the result of an outcross upon their inclusion in the collection, as Hulse-Kemp et al. (2015) explained when surveying the cotton germplasm collection. Population Five includes a subtree of the entire PD pedigree centered around the cultivar PD-3, all six of its descendants included in this study and two of its ancestors, and PD 6992, an outlier for this group with a low probability of true membership (43.9%). Population Six was the most diverse group, including germplasm releases resulting from crosses with elite materials from Delta Experiment Station, McNair, Deltapine, Stoneville breeding programs, and a line developed in China, ‘Jimian-8’ (May 1999). Fifty-nine of the 114 genotypes could not be classified into a single population at a probability ≥80%, providing evidence for the existence of significant admixture between groups.

Finally, an unrooted phylogenetic tree was generated using the general time reversible nucleotide substitution model to evaluate gene flow across the breeding program (**Figure 2.5**). The ability to resolve branches was fairly low, and most branches collapsed into polytomies due to low (<50%) bootstrap support, except for in cases with simple, unidirectional breeding schemes with noncyclic pedigrees. For example, unique clades containing the majority of fastSTRUCTURE Populations one, three, and four are obvious. This provides further evidence that even with a relatively small number of SNPs (~4500), we were able to draw insights about the history of breeding efforts in the PD program. Within-clade genetic variation was still relatively high, with branch lengths (proportional to genetic distance) > 0.1 usually present between sister taxa, indicating that gene flow across generations has contributed to the construction of multiple (10 clades with >3 member taxa), small (each clade < 20), diverse populations within the entire breeding program.

Following our analysis of the genetic variation with the PD germplasm, we identified genomic segments that distinguished PD genotypes from other improved *G. hirsutum* cultivars and breeding lines. Generally, PD genotypes tended to cluster together based on pairwise genetic distance (**Supplemental Figure 2.5**). For SNP loci passing filtering (CR > 90%, MAF > 2.5%), 3.5% of alleles were absent entirely from surveyed PD genotypes despite being present in the other improved *G. hirsutum* cultivars and breeding lines, whereas only 0.05% were private to the PD program, indicating that most of the SNP diversity present in the improved Upland cotton gene pool can be found in the PD program as well. Thirty-five putative selection windows were identified across 14 chromosomes, ranging from a single SNP with non-significant SNPs 25bp away to a larger region spanning 291kb in length, and these concentrated in the telomeric regions of each respective chromosome (**Figure 2.6**). Most of the SNPs under selection at (overall) minor allele at high frequency (~50%) in the PD genotypes and low frequency (< 5%) in the other improved *G. hirsutum* cultivars and breeding lines. Minor alleles for each of the 35 significant SNPs ($p < 0.05$) were present in every PD breeding group with low preference towards one breeding group over the others. Therefore, these chromosomal segments may be associated with the genetic background of the PD genotypes, regional adaptation, or the cumulative results of efforts to improve fiber quality traits, especially fiber strength (Campbell et al. 2011; Harrell 1974).

We further explored these regions by subjecting the genes in the putative selection window to gene enrichment analysis using gene ontology (GO) biological process annotations. We identified ten significant GO terms (Fisher's Exact Test $p < 0.05$) in five chromosomal regions associated with four categories of biological function: 1) response to auxin, 2) glutathione metabolic process, 3) actin nucleation, 4) and cellular localization and translation.

The four genes in the enrichment set annotated with the GO term "response to stimulus localized to a single 50kb in a segment of chromosome D02 (near 71.394 Mb). Although the role of auxin is ubiquitous across an array of morphological and immunological traits in plants, other genes in this enrichment set may give us a clue of how the PD programs breeding history has changed allele frequency in these particular regions. Gene expression studies in multiple plant species have exposed the potential for crosstalk between auxin biochemical pathways and other biotic and abiotic stress pathways (Lekshmy et al. 2017). These four genes are annotated as auxin-responsive protein small auxin up RNA (SAUR)-like, coding for small polypeptides (~140 amino acids) with an auxin-inducible motif. Other members of the SAUR gene family colocalize with fiber length and strength QTL (Li et al. 2017), and an association with fiber strength has been found nearby on D02 [qFS-Chr14-1.E1.XZV-RIL - (Shang et al. 2016)]. The minor alleles for these SNPs are found at about 40% frequency across breeding groups and is at <5% frequency in other improved cotton germplasm.

Two adjacent genes on chromosome D03 (6.39 - 6.40 Mb) were targets identified as gene set enrichment of glutathione metabolic process. These two genes (*D03G045000* and *D03G045100*) have not been previously identified as having a specific role in any gene pathways. The minor alleles at the nearby significant SNP was more prevalent in the earlier breeding groups than later breeding groups, suggesting a role in early germplasm development. Genes in the glutathione metabolic pathway in cotton have been found to associate with resistance to wilt caused by *Verticillium dahliae* and mediate salt stress (Li et al. 2019; Meloni et al. 2003).

A pair of tandem-repeat "formin-like protein 20" genes, annotated with the GO term "actin nucleation," were located near a significant SNP on chr A11 (3.35 Mb). Genes that affect the actin network that forms the cellular skeleton have been characterized as expressing in cotton fiber development and elongation (Li et al. 2005), and another gene that influences the actin network in cotton has been located in a selective sweep during domestication (Fang et al. 2017). Further work is needed to identify genes that influence cotton fiber formation and to determine if this locus is important for fiber production.

Five genes with the GO term "intracellular transport" and eleven with "translation" were also identified on chromosomes A11, D02, D03, and D09. Most of these genes have not been well characterized in cotton, although a few seem to be involved with host plant resistance. Seven of the eleven "translation" genes were annotated as involved in the "ribosome" pathway. One of the genes, *A11G030881*, a homolog of the *Arabidopsis ERF1* gene has been found to play a role in resistance to Verticillium wilt (Xu et al. 2011). One of the "intracellular transport" genes, *A11G032100*, is annotated as "vesicle transport v-SNARE 11-like", a member of family of genes that controls the transport of precursor molecules during gossypol production(Lang and Jahn 2008; Ting 2014). Gossypol levels are under genetic control and are thought to play a role in cotton host plant insect resistance (Liu et al. 2015).

We found evidence for sustained genetic diversity throughout eight breeding cycles of the PD program. Genetic signatures demarcating shifting breeding goals were evident after controlling for variable marker density across the genome. We also found SNP alleles with increased frequency in the PD program relative to in other improved upland cotton germplasm, with nearby genes enriched for biological functions including response to auxin, glutathione biosynthesis, translation, and cellular localization, implicating genetic drift for QTLs underlying host plant resistance. An additional locus under selection was found for actin nucleation, which may be a site that participated in fiber improvement in the Pee Dee program. The results of this study contribute to the growing body of knowledge regarding the breeding history of upland cotton in the southeastern US and the world. In addition, our findings in this study inform future breeding efforts based on PD program materials by establishing the basis for ongoing development of marker-assisted selection and genomic selection. The PD cotton germplasm enhancement program, an 85+ year old cotton improvement experiment, serves as a model system to study population genetics in the context of continued cotton improvement over the course of multiple breeders, breeding goals, and sources of genetic material.

References

Abdelraheem, A., H. Elassbli, Y. Zhu, V. Kuraparthy, L. Hinze *et al.*, 2020 A genome-wide association study uncovers consistent quantitative trait loci for resistance to Verticillium wilt and Fusarium wilt race 4 in the US Upland cotton. *Theoretical and Applied Genetics* **133** (2):563-577.

Ahn, S. J., J. Costa, and J. R. Emanuel, 1996 PicoGreen quantitation of DNA: effective evaluation of samples pre- or post-PCR. *Nucleic Acids Research* **24** (13):2623-2625.

Alexa, A., and J. Rahenfuhrer, 2020 topGO: Enrichment Analysis for Gene Ontology.

Beasley, J. O., 1940 The Origin of American Tetraploid Gossypium Species. *American Naturalist* **74** (752):285-286.

Bowman, A. W., and A. Azzalini, 2018 R package 'sm': nonparametric smoothing methods (version 2.2-5.6).

Bowman, D. T., and O. A. Gutierrez, 2003 Sources of Fiber Strenth in the U.S. Upland Cotton Crop from 1980 to 2000. *Journal of Cotton Science* **7**:164-169.

Calhoun, D. S., D. T. Bowman, and O. L. May, 1997 Pedigrees of Upland and Pima Cotton Cultivars Released Between 1970 and 1995, edited by M.A.F.E. Station.

Campbell, B. T., V. E. Williams, and W. Park, 2009 Using molecular markers and field performance data to characterize the Pee Dee cotton germplasm resources. *Euphytica* **169** (3):285-301.

Campbell, B. T., P. W. Chee, E. Lubbers, D. T. Bowman, W. R. Meredith *et al.*, 2011 Genetic Improvement of the Pee Dee Cotton Germplasm Collection following Seventy Years of Plant Breeding. *Crop Science* **51** (3):955-968.

Chang, C. C., C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell *et al.*, 2015 Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**:7.

Chen, Z. J., A. Sreedasyam, A. Ando, Q. Song, L. M. De Santiago *et al.*, 2020 Genomic diversifications of five *Gossypium* allopolyploid species and their impact on cotton improvement. *Nature Genetics* **52** (5):525-533.

Coop, G., D. Witonsky, A. Di Rienzo, and J. K. Pritchard, 2010 Using environmental correlations to identify loci underlying local adaptation. *Genetics* **185** (4):1411-1423.

Culp, T. W., D. C. Harrell, and T. Kerr, 1979 Some Genetic Implications in the Transfer of High Fiber Strength Genes to Upland Cotton. *Crop Science* **19** (4):481-484.

Culp, T. W., 1998 Public Breeding in the Southeast, pp. 493-519 in *Beltwide Cotton Conference*. National Cotton Council.

Deperi, S. I., M. E. Tagliotti, M. C. Bedogni, N. C. Manrique-Carpintero, J. Coombs *et al.*, 2018 Discriminant analysis of principal components and pedigree assessment of genetic diversity and population structure in a tetraploid potato panel using SNPs. *PloS One* **13** (3):e0194398.

Fang, L., Q. Wang, Y. Hu, Y. Jia, J. Chen *et al.*, 2017 Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits. *Nature Genetics* **49** (7):1089-1098.

Gapare, W., W. Conaty, Q.-H. Zhu, S. Liu, W. Stiller *et al.*, 2017 Genome-wide association study of yield components and fibre quality traits in a cotton germplasm diversity panel. *Euphytica* **213** (3).

Gauch, H. G. J., S. Qian, H. P. Piepho, L. Zhou, and R. Chen, 2019 Consequences of PCA graphs, SNP codings, and PCA variants for elucidating population structure. *PloS One* **14** (6):e0218306.

Goodstein, D. M., S. Shu, R. Howson, R. Neupane, R. D. Hayes *et al.*, 2012 Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research* **40** (Database issue):D1178-1186.

Gosselin, T., M. Lamothe, F. Devloo-Delva, and P. Grewe, 2020 radiator: RADseq Data Exploration, Manipulation and Visualization using R.

Gunther, T., and G. Coop, 2013 Robust identification of local adaptation from allele frequencies. *Genetics* **195** (1):205-220.

Hamblin, M. T., E. S. Buckler, and J. L. Jannink, 2011 Population genetics of genomics-based crop improvement methods. *Trends in Genetics* **27** (3):98-106.

Harrell, D. C., 1974 ARS-S-30: Breeding Quality Cotton and the Pee Dee Experiment Station Florence S.C., edited by USDA.

Hinze, L. L., A. M. Hulse-Kemp, I. W. Wilson, Q. H. Zhu, D. J. Llewellyn *et al.*, 2017 Diversity analysis of cotton (*Gossypium hirsutum* L.) germplasm using the CottonSNP63K Array. *BMC Plant Biology* **17** (1):37.

Huang, B. E., K. L. Verbyla, A. P. Verbyla, C. Raghavan, V. K. Singh *et al.*, 2015 MAGIC populations in crops: current status and future prospects. *Theoretical and Applied Genetics* **128** (6):999-1017.

Hulse-Kemp, A. M., J. Lemm, J. Plieske, H. Ashrafi, R. Buyyarapu *et al.*, 2015 Development of a 63K SNP Array for Cotton and High-Density Mapping of Intraspecific and Interspecific Populations of *Gossypium* spp. *G3 (Bethesda)* **5** (6):1187-1209.

Jombart, T., S. Devillard, and F. Balloux, 2010 Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics* **11**:94.

Kim, B., W. D. Beavis, and J. Leon, 2016 Numericware N: Numerator Relationship Matrix Calculator. *Journal of Heredity* **107** (7):686-690.

Klessig, D. F., M. Tian, and H. W. Choi, 2016 Multiple Targets of Salicylic Acid and Its Derivatives in Plants and Animals. *Frontiers in Immunology* **7**:206.

Kumar, S., G. Stecher, M. Li, C. Knyaz, and K. Tamura, 2018 MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular Biology and Evolution* **35** (6):1547-1549.

Lang, T., and R. Jahn, 2008 Core Proteins of the Secretory Machinery, pp. 107-127 in *Pharmacology of Neorotransmitter Release*, edited by T.C. Sudhof and K. Starke. Springer-Verlag Berlin Heidelberg.

Lekshmy, S., G. K. Krishna, S. K. Jha, and R. K. Sairam, 2017 Mechanism of Auxin Mediated Stress Signaling in Plants in *Mechanism of Plant Hormone Signaling under Stress*, edited by G. Pandey. John Wiley & Sons, Inc.

Li, X., G. Liu, Y. Geng, M. Wu, W. Pei *et al.*, 2017 A genome-wide analysis of the small auxin-up RNA (SAUR) gene family in cotton. *BMC Genomics* **18** (1):815.

Li, X. B., X. P. Fan, X. L. Wang, L. Cai, and W. C. Yang, 2005 The cotton *ACTIN1* gene is functionally expressed in fibers and participates in fiber elongation. *Plant Cell* **17** (3):859-875.

Li, Z. K., B. Chen, X. X. Li, J. P. Wang, Y. Zhang *et al.*, 2019 A newly identified cluster of glutathione S-transferase genes provides Verticillium wilt resistance in cotton. *Plant Journal* **98** (2):213-227.

Lischer, H. E., and L. Excoffier, 2012 PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* **28** (2):298-299.

Liu, X., B. Zhao, H. J. Zheng, Y. Hu, G. Lu *et al.*, 2015 Gossypium barbadense genome sequence provides insight into the evolution of extra-long staple fiber and specialized metabolites. *Scientific Reports* **5**:14139.

Maurer, A., V. Draba, Y. Jiang, F. Schnaithmann, R. Sharma *et al.*, 2015 Modelling the genetic architecture of flowering time control in barley through nested association mapping. *BMC Genomics* **16**:290.

May, O. L., 1999 Registration of PD 94042 Germplasm Line of Upland Cotton with High Yield and Fiber Maturity. *Crop Science* **39** (2):597-598.

Meloni, D. A., M. A. Oliva, C. A. Martinez, and J. Cambraia, 2003 Photosynthesis and activity of superoxide dismutase, peroxidase and glutathione reductase in cotton under salt stress. *Environmental and Experimental Botany* **49** (1):69-76.

Novembre, J., and M. Stephens, 2008 Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics* **40** (5):646-649.

Odong, T. L., J. van Heerwaarden, J. Jansen, T. J. van Hintum, and F. A. van Eeuwijk, 2011 Determination of genetic structure of germplasm collections: are traditional hierarchical clustering methods appropriate for molecular marker data? *Theoretical and Applied Genetics* **123** (2):195-205.

Paradis, E., and K. Schliep, 2019 ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35** (3):526-528.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick *et al.*, 2006 Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38** (8):904-909.

Raj, A., M. Stephens, and J. K. Pritchard, 2014 fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* **197** (2):573-589.

Shang, L., Y. Wang, X. Wang, F. Liu, A. Abduweli *et al.*, 2016 Genetic Analysis and QTL Detection on Fiber Traits Using Two Recombinant Inbred Lines and Their Backcross Populations in Upland Cotton. *G3 (Bethesda)* **6** (9):2717-2724.

Shen, C., X. Li, R. Zhang, and Z. Lin, 2017 Genome-wide recombination rate variation in a recombination map of cotton. *PloS One* **12** (11):e0188682.

Sun, Z., H. Li, Y. Zhang, Z. Li, H. Ke *et al.*, 2018 Identification of SNPs and Candidate Genes Associated With Salt Tolerance at the Seedling Stage in Cotton (*Gossypium hirsutum* L.). *Frontiers in Plant Science* **9**:1011.

Ting, H. M., 2014 Biosynthesis and transport of terpenes (Doctoral dissertation). Graduate School of Experimental Plant Sciences, Wageningen University, Wageningen, Netherlands.

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

USDA-ARS, 2015 Germplasm Resource Information Network (GRIN). in *Ag Data Commons.*, edited by U.A.R. Service, Beltsville, MD, USA.

Xu, L., L. Zhu, L. Tu, X. Guo, L. Long *et al.*, 2011 Differential Gene Expression in Cotton Defence Response to *Verticillium dahliae* by SSH. *Journal of Phytopathology* **159** (9):606-615.

Yang, Z., X. Ge, Z. Yang, W. Qin, G. Sun *et al.*, 2019 Extensive intraspecific gene order and gene structural variations in upland cotton cultivars. *Nature Communications* **10** (1):2989.

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

Figures and Tables

Figure 2.1. The historical relationships between Pee Dee breeding groups. The first four groups share a common gene pool primarily established in the first two breeding groups and focused on the improvement of fiber and agronomic characteristics. Groups five and six, focused on the development of host plant insect resistant breeding material and saw the introduction of new genetic diversity and background incorporated from group three. Groups seven and eight were formed from the combination of older, high quality material from the first four groups and new elite upland cultivars released from other breeding programs.

other breeding programs.

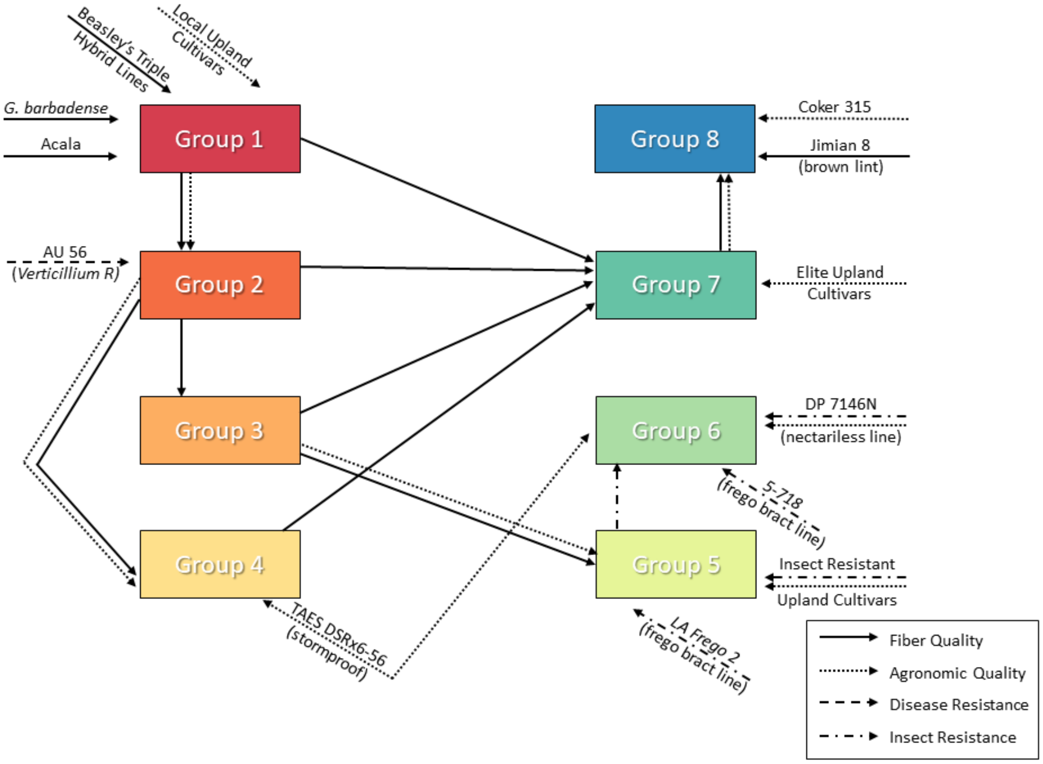


Table 2.1. Identity-by-state genetic distance for between- and within-breeding group comparisons, corrected for variable marker density. A higher number indicates that the individuals compared are more similar to each other, and lower numbers indicate individuals between groups are more different.

		Breeding Group							
Breeding Group		1	2	3	4	5	6	7	8
	1	0.673	0.680	0.669	0.667	0.636	0.633	0.650	0.646
	2	--	0.687	0.672	0.671	0.637	0.636	0.647	0.642
	3	--	--	0.686	0.689	0.663	0.659	0.668	0.662
	4	--	--	--	0.702	0.665	0.663	0.679	0.672
	5	--	--	--	--	0.682	0.698	0.662	0.658
	6	--	--	--	--	--	0.713	0.659	0.657
	7	--	--	--	--	--	--	0.685	0.672
	8	--	--	--	--	--	--	--	0.676

Figure 2.2. Comparison between two principal component estimation methods before and after correcting for variable marker density. The SNP x Individual biplots of the principal component (PC) coordinates for individuals, colored by breeding group, in PC1 (horizontal axis) and PC2 (vertical axis). Percent variance explained by each of the first two PCs was calculated by dividing the eigenvalue of the PC by the sum of the eigenvalues for the first 40 PCs. A plink PCA with 17,441 SNPs, B double-centered PCA with 17,441 SNPs, C plink PCA with 4,597 SNPs out of strong LD (R2 < 0.8) double-centered PCA with 4,597 SNPs out of strong LD (R2 < 0.8).

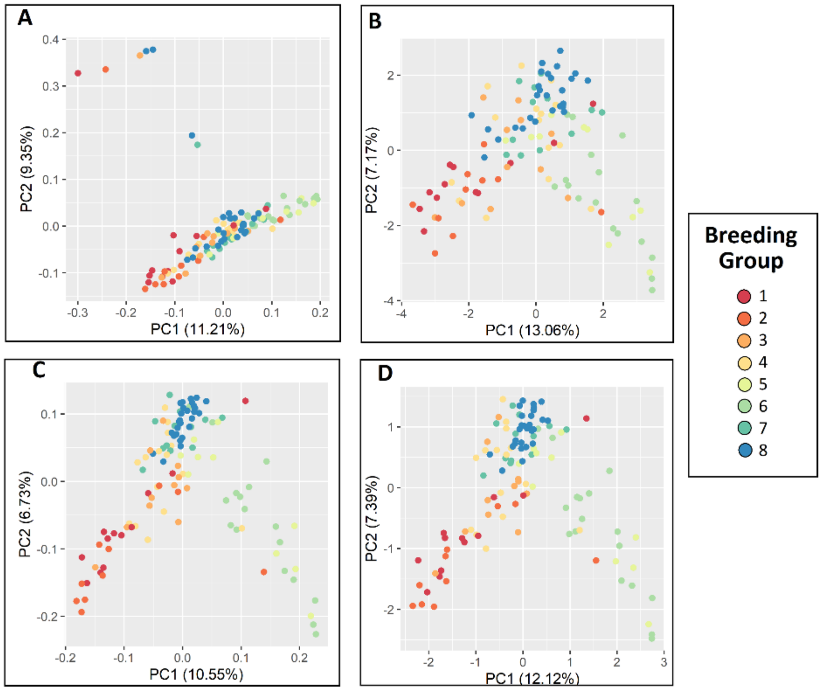


Figure 2.3 The Q plot for six fastSTRUCTURE subpopulations. Membership probability plot for probability of group assignment, sorted by the likeliest group assignment for each individual. The most likely number of populations (k), as determined by the model complexity that maximizes marginal likelihood, is 6. The individual names are given along the bottom of the horizontal axis, with the breeding group number given above it in the same color scheme as other figures.

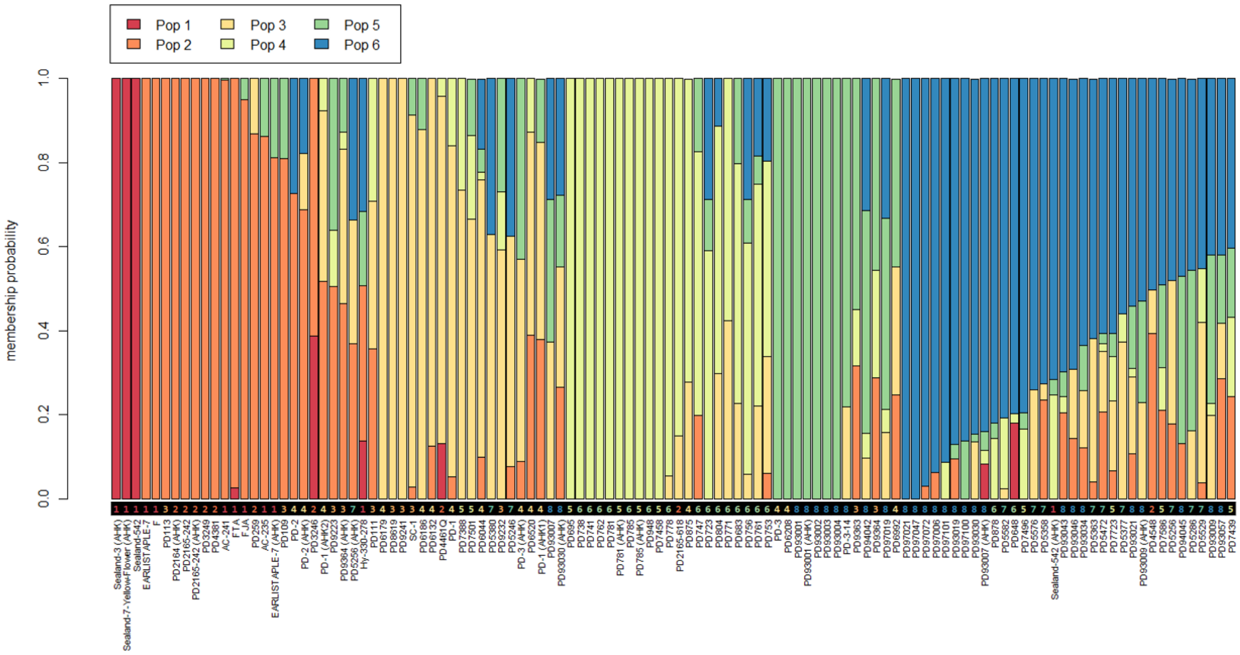


Figure 2.4. Overlap between three group designation methods. Sankey diagram showing how individuals in each of the prior breeding groups (center) are classified in fastSTRUCTURE (left) and in discriminant analysis of principal components (DAPC) (right). In both DAPC and fastSTRUCTURE, the number of populations or clusters (*k* = 6 for fastSTRUCTURE, *k* = 3 for DAPC) is less than the number of breeding groups (*k* = 8).

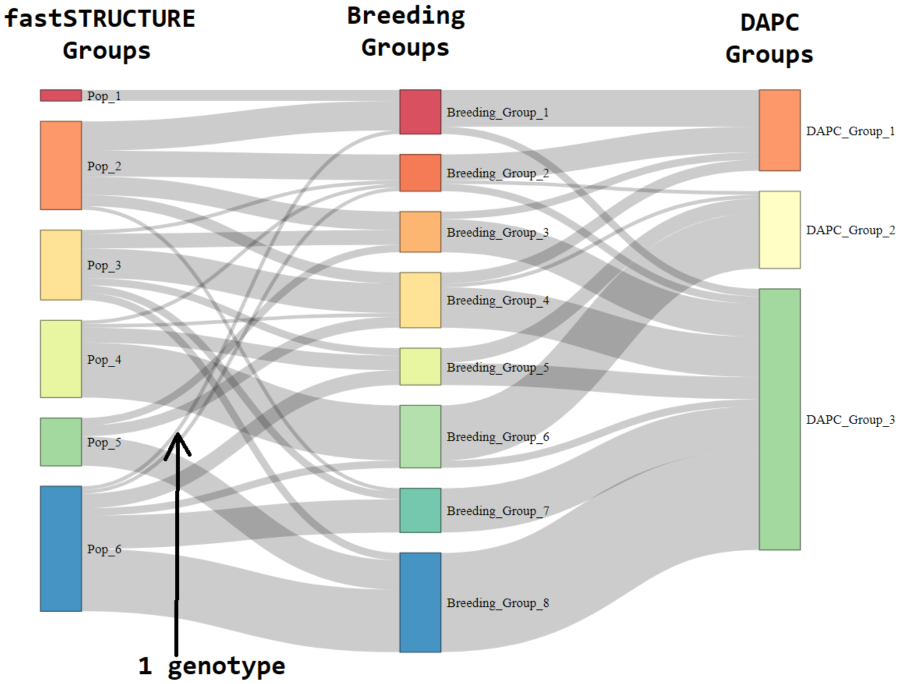


Figure 2.5. Unrooted consensus phylogenetic Tree for 114 Pee Dee genotypes. Phylogenetic analysis was performed in MEGA X with the general time reversible model (G=3 classes of evolutionary rates). Bootstrap values are given for branches with >50% support based on 1000 replicates, and other branches are collapsed into polytomies. Branch length is proportional to the genetic distance between sub-branches. Unresolved nodes are expected due to high admixture and inbreeding across breeding generations. Highlighted clades correspond to populations discovered with fastSTRCTURE.

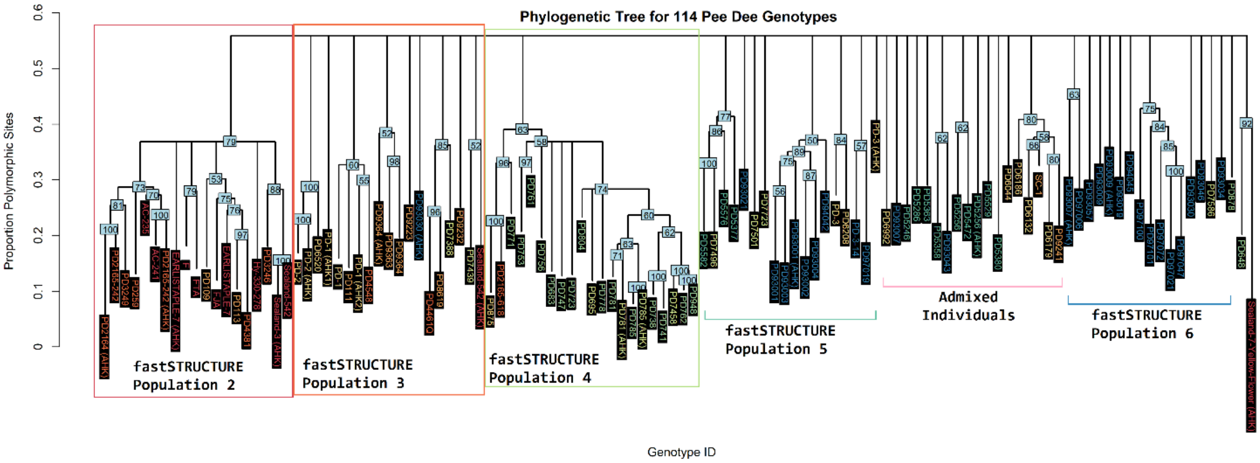

Phylogenetic Tree for 114 Pee Dee Genotypes
Proportion Polymorphic Sites
0
0.1
0.2
0.3
0.4
0.5
0.6
Genotype ID
fastSTRUCTURE Population 2
fastSTRUCTURE Population 3
fastSTRUCTURE Population 4
fastSTRUCTURE Population 5
Admixed Individuals
fastSTRUCTURE Population 6
PD2164 (AHK)
PD-3 (AHK)
PD648
Seabrook Yellow Flower (AHK)

Figure 2.6. Identifying loci under selection in the Pee Dee Breeding Program. A The log10 Bayes Factor from BayEnv for each of the 20,566 SNPs that were significant in separating out the114 Pee Dee from the 249 other improved Upland cotton genotypes. B Allele frequency for six significant SNPs in Pee Dee breeding groups one through eight (1-8) or other genotypes (W) are given on the horizontal axis. The red numbers in A and B indicate significant SNPs that are near genes annotated with significant gene ontology terms.

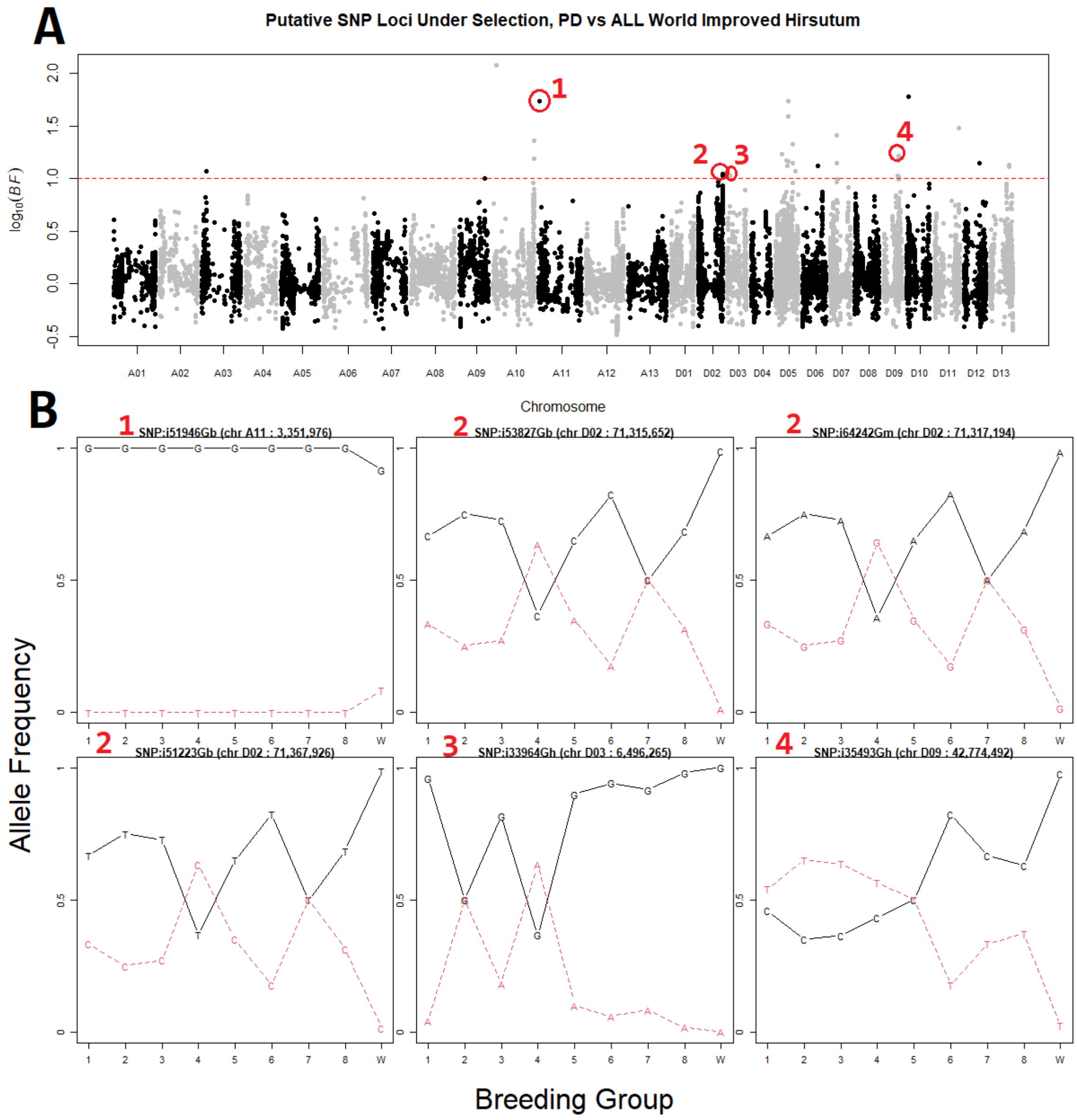


Table 2.2. Significant Gene Ontology : Biological Process Terms in regions under selection. Gene ontology (GO) terms for Biological Process enriched in the set of genes in genomic regions (detected with BayEnv) that differentiate Pee Dee genotypes (n=114) from other improved worldwide G. hirsutum material (n=249) filtered to include only those terms significant by the graph weight method for Fisher's exact test (p-value < 0.05).

		Number of Genes with this GO Term				**p-value**	
GO Number (Biological Process)	**Gene Ontology Term (1,341 terms > 5 genes)**	**Count in Whole Genome (*n*=24,647 genes with GO annotation)**	**Count in Selection Windows (*n*=52 genes with GO annotation)**	**Expected (of 52 randomly chosen genes)**	**Fisher's Exact Test Rank**	**Weight Method**	**Fisher's Exact Test**
GO:0006749	glutathione metabolic process	9	2	0.02	3	0.00016	0.00016
GO:0006412	translation	1494	11	3.15	4	0.00024	0.00075
GO:0009733	response to auxin	263	4	0.55	11	0.00230	0.00230
GO:0046907	intracellular transport	489	5	1.03	14	0.00364	0.00364
GO:0045010	actin nucleation	44	2	0.09	16	0.00390	0.00390
GO:0044743	protein transmembrane import into intracellular organelle	18	1	0.04	67	0.03732	0.03732
GO:0006452	translational frameshifting	20	1	0.04	69	0.04138	0.04138
GO:0009416	response to light stimulus	20	1	0.04	70	0.04138	0.04138
GO:0045901	positive regulation of translational elongation	20	1	0.04	71	0.04138	0.04138
GO:0045905	positive regulation of translational termination	20	1	0.04	72	0.04138	0.04138

Supplemental Methods

Anchoring Marker Probe Sequences to Reference Genome

Complete marker flanking sequences were downloaded from Hulse-Kemp et al. (2015). The strand orientation was flipped to match the strand indicated in the project file Illumina Genome Studio. The 50 nucleotide sequence upstream of each probe sequence was extracted and saved into a fasta file, with each sequence labeled as the corresponding project marker name. The v2.0 *Ghir* reference genome assembly (Chen et al. 2020) was downloaded from Phytozome (Goodstein et al. 2012). A local BLAST database was built with the "makeblastdb" command. The probe sequences were queried against the database with the "blastn" command. The strict set of matching BLAST hits were filtered to only those with a minimum match length of 45 or longer. A more lenient set was generated to include lower e-value matches with another run of "blastn."

A custom python script was used to combine information from the F_2 intraspecific genetic map presented in Hulse-Kemp et al. (2015), inter-marker correlations, and BLAST hits. First, reciprocal best matches were identified based on inter-marker correlation, such that pair of highly correlated markers were identified ($R^2 > 0.8$). The markers were anchored to the reference genome if the highest e-value BLAST hits for both markers were within 5 Mb on the same chromosome. The markers were not anchored if the chromosome assignment disagreed with the linkage group assignment from the F_2 map. A random subset of 20% of the already anchored markers were chosen to extend the number of anchored markers to those with high inter-marker correlation with an already anchored marker, further choosing the most likely BLAST hit between high quality choices. Next, the remaining markers with F_2 map positions were allocated to the corresponding pseudomolecule and inserted only if there was at least one nearby marker already inserted that was correlated with that marker.

This left a few types of markers: 1) those with a disagreement between the lowest e-value BLAST hit and the chromosome assigned from the F2 map, 2) markers absent on the genetic map with competing best insertion positions based on inter-marker correlation and lowest e-value BLAST hit, and 3) markers that either lacked a high quality BLAST hit or were not highly correlated with a nearby marker. To identify the best fitting insertion point for each marker, a random marker was chosen repeatedly until all markers had been addressed. For each marker, a goodness of fit score was assigned to each BLAST hit, providing a better score to insertion points with anchored markers with high inter-marker correlations with the selected marker. The score was calculated as the sum product of pairwise R^2 and 1/log10(distance between BLAST hit and anchored marker + 10). At first, only those markers with the lowest e-value BLAST hit and LD-based score were inserted until no more markers could be anchored. Accordingly, tie-breaking was enabled, which showed a preference to the LD score over the BLAST hit e-value. Once tie-breaking yielded no further anchored markers for markers that either had no good BLAST hits or had no correlation with already anchored markers, the low-quality BLAST hits were evaluated instead.

This process was repeated 1000 times for various thresholds of inter-marker correlations, chosen from a uniform distribution ranging from R^2=0.2-0.79. The results from bootstrapping were filtered to include markers that were successfully anchored to any chromosome in at least 80% of trials and mapped at least 20% more to the most frequent choice than the second most frequent choice.

Marker Density

To explore changes in the distribution of SNP marker loci across potential MAF values, between 0.025 and .500, the “density.compare” function in the R package sm was used (Bowman and Azzalini 2018). The nonparametric test for density equality, using the “model=equal” flag, was also performed, using the optimal density parameter, *h*. Next, the “sm.density.compare” function was used to evaluate changes in SNP marker density across chromosomes for the mapped markers in datasets one and two. The same nonparametric test for density equality was used (* indicates $p < 0.05$).

Fit Against Pedigree Data

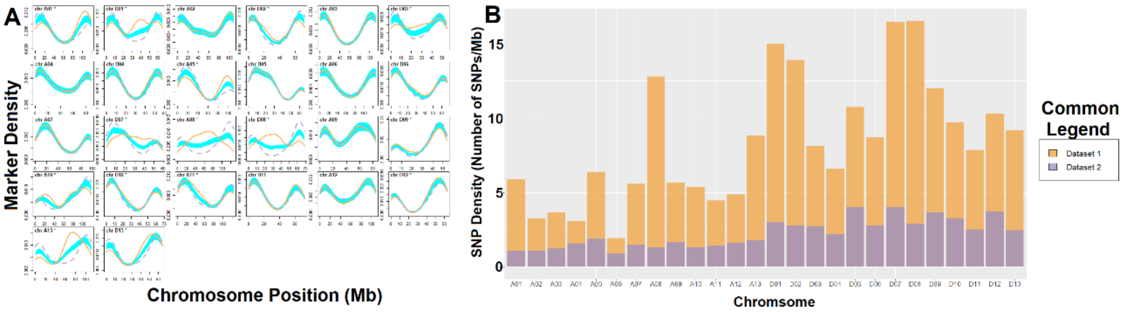
The pairwise identity by state (IBS) genetic distance matrix was generated in plink 1.9 with the “--dist 1-ibs” command for datasets one and two. Expanded pedigrees were used to calculate the generalized numerator relationship matrix, a value proportional to the expected percentage of identity by descent (IBD) alleles between individuals, with the NumericwareN software (Kim et al. 2016). Goodness of fit between IBS measurements of the two datasets was estimated by plotting the two against each other, and regression statistics calculated using “lm” with the formula “plink_IBS_dist_dataset2 ~ plink_IBS_dist_dataset1.” To test for goodness of fit to each of the pairwise genetic matrices, regression analysis was performed on the observed SNP-based IBS genetic distance for datasets one and two as explained by expected IBD estimate for each pair of genotypes. Regression statistics were calculated using the “lm” function in R, with the formula “plink_IBS_dist_datasetn ~ NumericwareN_IBD.”

Supplemental Figures

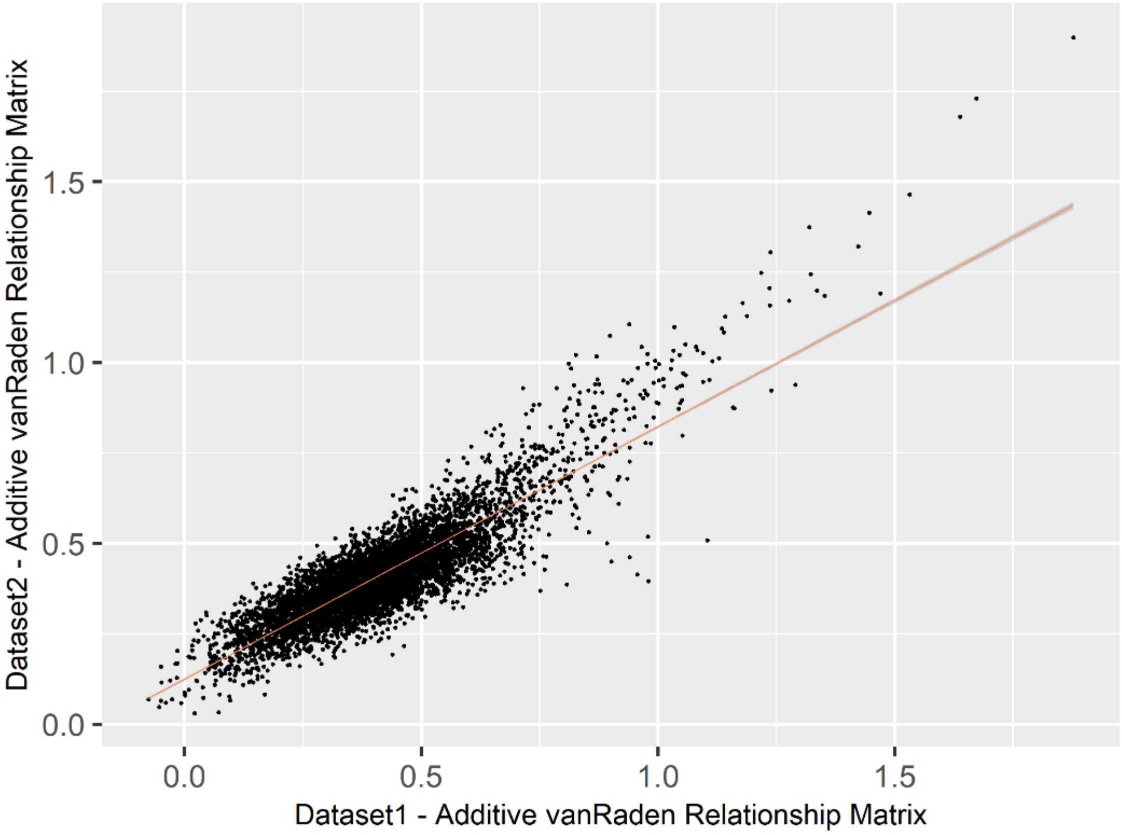
Supplemental Table 2.1. Non-default settings for software used in Chapter Two. The explanation for the settings is also given.

Task	Program	Command	Flags/Options	Explanation
n/a	plink v1.9	ALL	--autosome-num 26	Sets the chromosome set to 26 chromosomes
			--allow-no-sex	Disables the no-sex warnings
Generating Thinned Data Set		--indep-pairwise	2500 kb	Set the window size to 2.5 Mb
			1	Set step-size at 1 marker, so all adjacent markers are tested
			0.8	Sets the R^2 threshold for considering SNPs to be independent
Discriminant Analysis of Principal Components	adegenet (R Package)	dapc	n.pca = 75	Set the initial number of principal components in model to 75
			n.da = 5	Calculate 5 discriminant functions in the first pass
Detecting markers under selection	BayEnv2	bayenv2	-i X	Input the file X.txt with the allele counts for the two groups. Each locus is run separately.
			-e envfile.txt	Enter in a dummy environmental variable file, set to -0.707 and 0.707 for the two groups
			-m mat.txt	File with the background likelihood of differentiation at a random SNP locus
			-k 100000	Perform 100,000 iterations in the Markov chain Monte Carlo.
			-r $RANDOM	Generates a random seed for each model run
			-p 2	Sets the total number of populations to 2 (PD and non-PD)
			-n 1	Sets the number of environments being tested to 1 (the dummy environment)
			-t	Runs in "test" mode, to generate Bayes factors for each SNP
			-s sizes.txt	Provides the population sizes for the two groups, to account for missing data
			-o X	Saves the output to file X (depends on which SNP file was input)
Aligning SNP array probes to reference genome	blast	makeblastdb	-dbtype = nucl	Sets the database type to nucleic acid
			-input_type = fasta	Input file type is FASTA formatted
			-paste-seqids	Includes the SeqIDs, which in this case are the chromosome/scaffold assemblies
	blast [strict matches]	blastn	-outfmt = 6	Return in a tab delimited format with blast style 6.
			-num_threads = 4	Use all four CPU threads
			-num_alignments = 10	Return the ten best alignments
			-perc_identity = 98	Only extract matches with a minimum identity of 98%
	blast [lower quality matches]	blastn	-outfmt = 6	Return in a tab delimited format with blast style 6.
			-num_threads = 4	Use all four CPU threads
			-num_alignments = 10	Return the ten best alignments
			-perc_identity = 98	Only extract matches with a minimum identity of 96%
			-word_size = 9	Use a shorter word size (oppose to normal word size 11) to allow for more alignment or errors

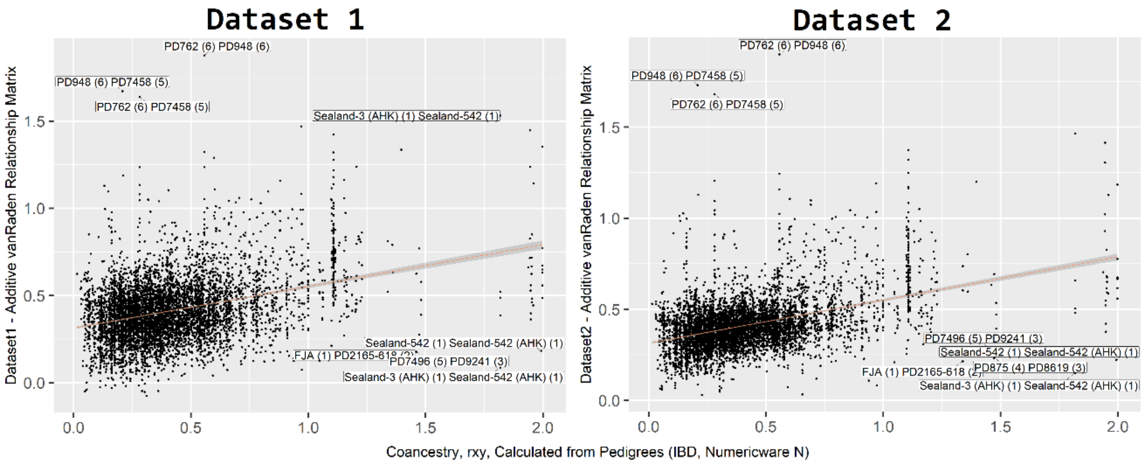
Supplemental Figure 2.1. .SNP density on each chromosome before and after removing redundant markers. A Marker density, given as the percentage of markers on that chromosome within h (~15 Mb for A sub-genome chromosomes, ~5 Mb for D), the optimal smoothing parameter, of a given position. The permutation test of equality was used to determine whether the collection of SNP markers could have come from the same underlying distribution, i.e. if they represent the same density of markers (markers/Mb in the smoothing window) along the entire chromosome. Chromosomes with significantly different marker densities (p < 0.05) are marked with a "*". B The marker density, normalized by chromosome size, before and after removing redundant markers.



Supplemental Figure 2.2. Pairwise additive relationship values for 6441 combinations of 114 genotypes in Dataset One and Two. The line of best fit (m = 0.70, b = 0.12, R2 = 0.77) is plotted in orange. Individuals that show high to moderate genetic differences tend to have the largest overall change in IBS between the two datasets.



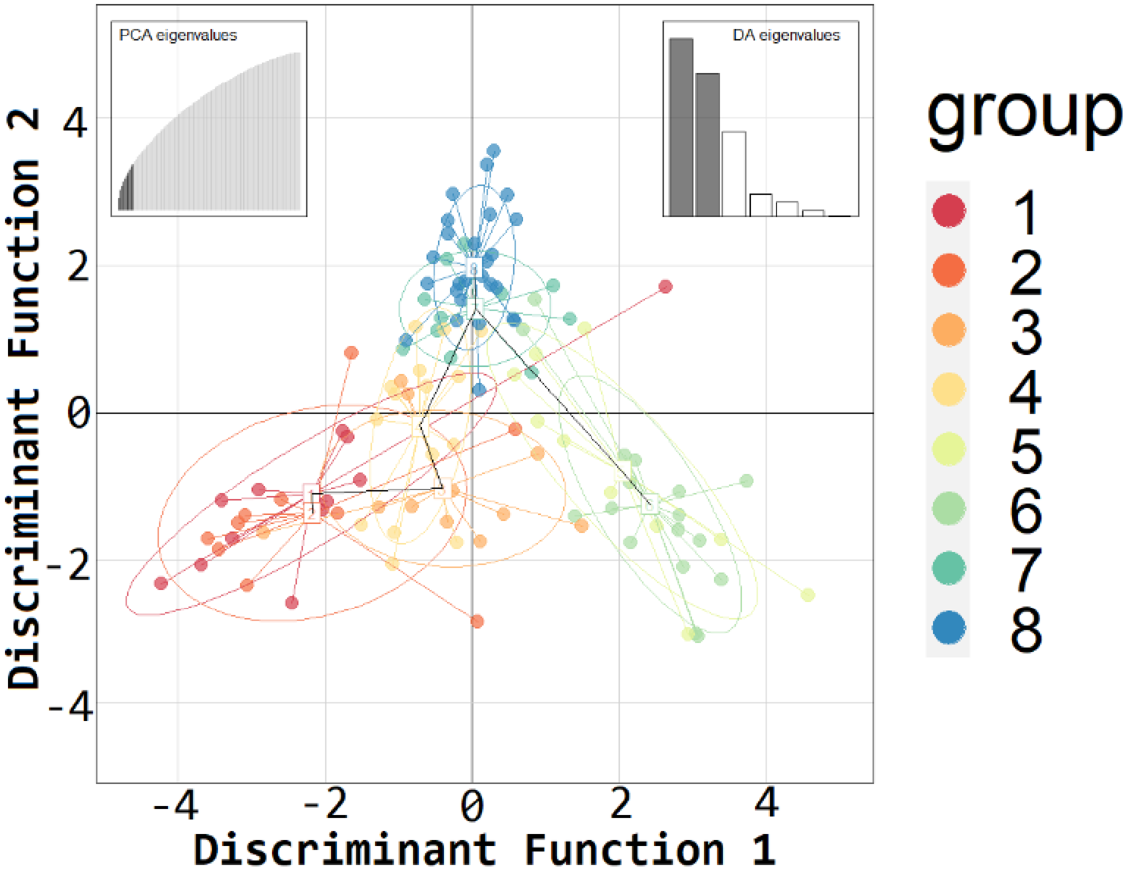
Supplemental Figure 2.3. Goodness of fit for pairwise genetic relatedness against pedigrees, before and after correcting for marker redundancy. Observed genetic relationship matrix for Dataset One (A) and Dataset Two (B) plotted against coancestry calculated from extended pedigrees in NumericwareN. The line of best fit for Dataset One (m = 0.19, b = 0.53, R^2 = 0.13) and Dataset Two(m = 0.07, b = 0.82, R^2 = 0.20) are given in orange. A stronger positive association is apparent in B, where the observed genetic distance values tend to align more with the expected value. The cluster of points around x = 1 in A and B is due to the large number of comparisons between full-sib genotypes. Outlier genotype pairs have the ID and breeding group number in call-outs.



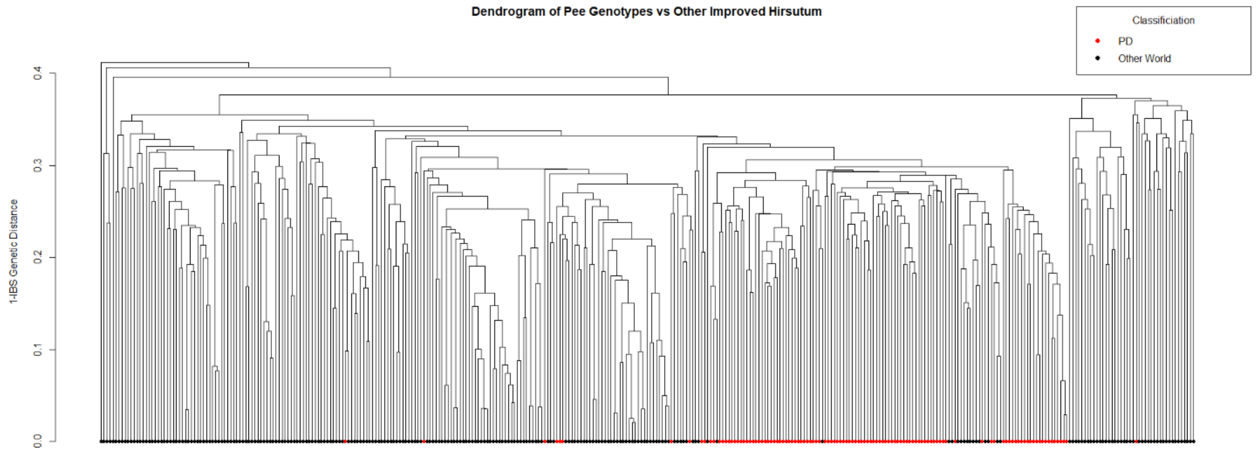
Supplemental Table 2.2. Heterozygosity for each genotype in Dataset One and Two.

Genotype	Percent of Markers that are Heterozygous		Rank	
	Dataset 1	Dataset 2	Dataset 1	Dataset 2
PD804	37.38%	35.75%	1	1
PD761	33.85%	32.32%	2	2
PD-3 (AHK)	32.18%	30.66%	3	3
PD94042	30.05%	29.81%	4	5
AC-235	27.68%	30.62%	5	4
PD5576	25.86%	19.77%	6	8
PD7723	21.20%	20.22%	7	7
PD5256	20.20%	17.47%	8	13
PD6044	20.15%	22.44%	9	6
PD5363	19.22%	16.62%	10	15
PD94045	19.02%	16.20%	11	17
PD93021	18.53%	18.61%	12	11
PD5472	17.99%	17.14%	13	14
Sealand-542	17.69%	18.85%	14	10
PD5358	17.49%	15.76%	15	18
PD5286	17.37%	14.25%	16	23
PD5529	17.14%	18.40%	17	12
PD93009	16.77%	16.61%	18	16
PD756	15.98%	14.40%	19	21
SC-1	15.54%	15.53%	20	19
PD6186	15.00%	18.88%	21	9
PD785 (AHK)	14.73%	12.60%	22	27
AC-241	14.56%	14.05%	23	24
PD-2	14.49%	14.27%	24	22
PD5246	14.06%	13.70%	25	26
PD785	13.93%	13.83%	26	25
PD753	13.44%	14.96%	27	20
PD771	12.74%	12.05%	28	28
PD6132	12.54%	11.58%	29	29
PD5256 (AHK)	11.26%	8.29%	30	38
PD93007 (AHK)	10.68%	9.39%	31	35
PD-1 (AHK1)	10.35%	10.31%	32	32
PD5377	9.89%	9.39%	33	34
PD7586	9.30%	11.29%	34	31
PD93009 (AHK)	9.17%	11.44%	35	30
PD97047	8.95%	4.63%	36	56
PD878	8.94%	10.12%	37	33
PD93001	8.86%	5.12%	38	54
PD4548	8.25%	8.23%	39	39
PD778	8.18%	8.54%	40	37
PD9363	8.05%	5.52%	41	50
PD741	7.85%	7.39%	42	41
PD97072	7.81%	9.19%	43	36
PD93019	7.74%	6.36%	44	47
PD97101	7.61%	6.77%	45	45
PD-2 (AHK)	7.55%	7.16%	46	43
PD9364	7.41%	7.65%	47	40
PD97100	7.14%	7.36%	48	42
PD93043	6.14%	5.71%	49	48
PD259	6.01%	4.66%	50	55
PD93004	5.54%	4.30%	51	59
PD648	5.46%	6.91%	52	44
PD7388	5.12%	4.35%	53	57
PD3246	5.08%	6.41%	54	46
PD93001 (AHK)	4.98%	5.57%	55	49
PD93034	4.42%	5.24%	56	53
PD7496	4.33%	5.49%	57	51
PD5582	4.13%	4.11%	58	62
PD9223	3.97%	4.22%	59	61
PD-1	3.82%	5.47%	60	52
PD93030	3.32%	3.33%	61	70
PD93057	3.05%	2.99%	62	73
PD111	3.01%	4.27%	63	60
PD-3-14	3.00%	2.70%	64	77
PD93007	2.99%	2.60%	65	80
PD6179	2.98%	3.71%	66	66
PD948	2.92%	3.11%	67	71
PD6992	2.84%	4.31%	68	58
PD9232	2.83%	3.88%	69	64
F	2.77%	3.86%	70	65
PD738	2.77%	3.64%	71	67
PD93002	2.76%	2.92%	72	75
Hy-330-278	2.65%	3.39%	73	69
PD97019	2.64%	4.10%	74	63
PD3249	2.45%	2.29%	75	86
PD747	2.40%	3.46%	76	68
PD683	2.31%	2.00%	77	90
PD93003	2.29%	3.05%	78	72
PD781	2.28%	2.81%	79	76
PD781 (AHK)	2.20%	2.98%	80	74
PD7458	2.18%	2.61%	81	79
PD723	2.08%	2.46%	82	84
PD2165-618	1.94%	1.92%	83	91
PD93046	1.90%	2.66%	84	78
PD9364 (AHK)	1.83%	2.51%	85	83
FTA	1.80%	2.42%	86	85
Sealand-542 (AHK)	1.78%	2.18%	87	88
PD109	1.75%	2.05%	88	89
PD97006	1.72%	2.55%	89	82
PD97021	1.70%	2.58%	90	81
PD93030 (AHK)	1.54%	2.29%	91	87
PD-1 (AHK2)	1.53%	1.49%	92	95
PD113	1.38%	1.92%	93	92
EARLISTAPLE-7	1.21%	1.76%	94	94
FJA	1.18%	1.48%	95	96
PD7439	1.15%	0.98%	96	103
EARLISTAPLE-7 (AHK)	1.15%	1.35%	97	98
PD5380	0.96%	1.48%	98	97
PD8619	0.93%	1.22%	99	101
PD-3	0.89%	1.81%	100	93
PD6520	0.87%	1.24%	101	100
PD6208	0.76%	1.24%	102	99
PD2165-242	0.58%	1.09%	103	102
PD4381	0.57%	0.74%	104	105
PD762	0.47%	0.76%	105	104
PD695	0.32%	0.41%	106	108
Sealand-7-Yellow-Flower (AHK)	0.32%	0.41%	107	107
PD9241	0.29%	0.50%	108	106
Sealand-3 (AHK	0.28%	0.33%	109	110
PD7501	0.26%	0.30%	110	111
PD2164 (AHK)	0.21%	0.26%	111	114
PD875	0.20%	0.28%	112	112
PD4461Q	0.19%	0.33%	113	109
PD2165-242 (AHK)	0.19%	0.28%	114	113

Supplemental Figure 2.4. Discriminant analysis of principal components for the eight Pee Dee breeding groups. The DAPC biplot for Dataset Two, with coordinates for each individual in discriminant function 1 (DF1, horizontal axis) plotted against discriminant function 2 (DF2, vertical axis). Individuals are represented by a point, color-coded for each breeding group. The ovals represent the expected spatial distribution of individuals in DF1 and DF2.



Supplemental Figure 2.5. Dendrogram of the maximum likelihood neighbor-joining tree for 363 cotton genotypes. The plot was generated through hierarchical clustering on the genetic distance matrix for114 Pee Dee and 249 other improved Upland cottons. Branch length is proportional to the genetic distance between the two child nodes. Pee Dee genotypes, the leaves labeled in red, tend to cluster together with a few outliers. Within-group genetic diversity is similar to genetic variation in other clades. That subtree topology is similar to that of just the 114 genotypes from this study.



CHAPTER THREE

GENETIC ARCHITECTURE OF COTTON AGRNOMIC PERFORMANCE AND FIBER QUALITY IN THE PEE DEE GERMPLASM ENAHANCEMENT PROGRAM

Abstract

The Pee Dee Cotton Germplasm Enhancement Program has developed improved upland cotton (*Gossypium hirsutum* L.) genotypes for the Coastal Plains region of the southeastern US for over 80 years. This closed breeding program contains extensive genetic variation for fiber quality traits, which has been utilized over the past few decades as a source of improved fiber strength and fiber length for public and commercial breeding efforts. An extensive genetic survey of the resources in the Pee Dee program was conducted using a combination of 17,226 filtered SNP markers with 14 year-locations (environments) of previously reported agronomic performance and fiber quality data. Thirty-three independently segregating haplotype blocks associated with variation for agronomic performance or fiber quality were identified using a kernel-based, mixed linear model for haplotype-set genome-wide association. Hierarchical clustering and haplotype binning revealed 16 previously unreported QTL. The strongest QTL signals were detected in a set of ten haplotype blocks across chromosome D06. These QTL for fiber length, strength, and gin turnout were detected across environments. SNP data revealed potential routes of gene flow to and from the Pee Dee program. The results of this study provide a basis for genomic selection strategies or pyramiding beneficial haplotypes.

Introduction

The farm gate value of US upland cotton (*Gossypium hirsutum* L.) exceeded $12 billion USD during 2019 (Johnson et al. 2020). Lint value is determined primarily by lint yield, but fiber quality is also important to meet the needs of textile manufacturers. Improved cultural practices optimize yield and fiber quality performance (Lewis et al. 2000; Viator et al. 2005; Bednarz et al. 2005); however, improved cultivars provide a baseline for productivity and are key to enhanced production (Bowman 2000). Cotton breeders often examine yield components such as seed index, bolls m^{-2}, gin turnout, and boll weight when evaluating lint yield (Meredith and Wells 1989; Jenkins et al. 1990; Lewis et al. 2000). Negative correlations among these traits makes selection for improved overall total lint yield challenging, especially in the context of conventional breeding (Tang et al. 1996; Campbell et al. 2012).

Cotton fiber quality traits are equally complex, and substantial research has been conducted to identify the genetic basis of fiber quality (Paterson et al. 2003; Fang et al. 2014; Li et al. 2016; Islam et al. 2016; Fang et al. 2017; Chandnani et al. 2018; Naoumkina et al. 2019; Thyssen et al. 2019). Fiber quality is most often measured using two machines including the high volume instrument (HVI) and the advanced fiber information system (AFIS). The HVI measures the characteristics of a bundle of fibers, whereas the AFIS measures individual fibers. Global fiber classification is performed with the HVI (Foulk et al. 2007). Textile mills value the length and strength of the fibers, the textural properties (micronaire, fineness, and maturity ratio), and the overall uniformity of the fibers (Foulk et al. 2007).

Varying heritability estimates for yield and fiber quality indicate the complex basis of these traits and the significant interplay between genotype and environment (Paterson et al. 2003; Khan et al. 2017; Campbell and Jones 2005). Genome-wide association studies have revealed part of the underlying architecture of multiple fiber quality and yield-related traits in upland cotton (Thyssen et al. 2019; Hinze et al. 2017; Du et al. 2018; Ma et al. 2018; Huang et al. 2018). These studies have used genetic markers (SNPs or SSRs) by fitting a linear model or mixed linear model (MLM) on a single marker at a time. Results from single marker analysis studies have a straightforward biological interpretation because an additive or dominance model is used to score the effect of a (minor) allele. Also, direct estimates of the effect of a single nucleotide polymorphism (SNP) or simple sequence repeat (SSR) can be calculated using regression analysis and the associated test statistics (Korte and Farlow 2013). These tests normally treat SNPs as fixed effects (Zhang et al. 2010), which can be tested for interactions with other model terms, although random effect models also exist (Wang et al. 2016). The disadvantage of single marker analysis is the inability to fully account for linkage disequilibrium (LD) structure, epistatic interactions between genes, or marker redundancy (Wang et al. 2011).

In structured populations, such as those from breeding programs or diversity panels, higher relatedness than expected among individuals (‘cryptic relatedness’) makes it difficult to avoid confounding due to common ancestry between genotypes (Astle and Balding 2009). Domesticated upland cotton is derived from a common gene pool, with successive rounds of sub-selection causing at least two identifiable genetic bottlenecks (Iqbal et al. 2001). Prior studies have demonstrated the ability of model covariates to efficiently correct for population substructure that can skew genome wide association studies (GWAS) results, such as principal component analysis and STRUCTURE subpopulation groupings (Price et al. 2006; Odong et al. 2011) and variance component partitioning through the decomposition of the kinship matrix (Sun et al. 2010). At the expense of statistical power, the practical consequence of population structure correction is the reduction in the discovery rate for significantly associated genetic markers (Shin and Lee 2015). Therefore, there is a balance between increasing model complexity and the ability to identify (non-)beneficial variants and simple practical applications, especially for plant breeders who are interested in better utilizing germplasm resources.

A key assumption in single marker analysis is that markers assort independently (Waksmunski et al. 2020), enabling one to evaluate the significance of each marker separately without considering nearby co-segregating markers. Composite interval mapping can help build association models that integrate nearby markers, although it is normally reserved for use in biparental populations with some recent exceptions (Wang et al. 2016). At the significant loss of information, one can also reduce (or thin) markers using local LD or a fixed window size to reduce redundancy and the downstream computational load (Li et al. 2018). Another recent advancement is the category of gene-set tests (variably called SNP-set, haplotype-set, etc.) that test multiple related markers together, usually with a correction for the number of markers tested (Wang et al. 2011).

By combining the principal of kernel based tests (Yang et al. 2008; Morota and Gianola 2014) with the concept of gene-sets, Wu et al. (2011) developed the sequencing kernel association test (SKAT), with further adaptation to the genetic MLM (with kinship) in the reliable association inference by optimizing weights (RAINBOW) method. Complex kernels are useful for detecting "hidden" signals that may be useful for genomic selection, but are difficult to interpret from a biological perspective (Morota and Gianola 2014). One of the simplest kernels is the linear kernel, which is derived by calculating the local additive genomic relationship matrix for that gene-set (VanRaden (2008). After the kernel is identified, eigen decomposition or another dimensional reduction technique can be used and the resulting model can be solved with efficient mixed model association (Kang et al. 2008) A *p*-value for each gene-set is produced using the likelihood ratio test or score test, which can be directly calculated by comparing the model with and without the kernel for that gene-set.

If haplotypes of co-segregating markers are binned as gene-sets, these haplotype-sets can be tested one at a time for association with a trait of interest. The linear kernel calculated for each haplotype-set is a transformation on the pairwise genetic distance for each combination of individual genotypes in the study, and the test statistic from RAINBOW reflects whether or not the individuals with similar haplotypes have similar phenotypic values. This is different than the test performed in the classic additive model with single marker analysis or a multi-locus model, where the effect of each genetic marker is directly estimated, and regression analysis is performed based on allele count. The haplotype-based kernel association method is fundamentally different, because the haplotype is treated as a random effect, and local similarity across multiple loci is what is driving the signal detection. Since haplotype association reduces the number of individual tests performed, multiple-test correction procedures can be relaxed, increasing power relative to single marker analysis, especially when detecting effects due to rare variants (Wu et al. 2011). Ideally, haplotype-set GWAS can enable the identification of favorable haplotypes in QTL regions, which can be used for plant breeding (Su et al. 2016).

For plant breeders, haplotype-set GWAS has many benefits as opposed to classical single marker analysis. Although marker assisted selection schemes have demonstrated success for simple, mendelian traits or those with genes of major effect (Fang et al. 2010; Chandnani et al. 2018; Abdelraheem et al. 2020), selection on haplotypes as a whole has the potential to capture some of the missing genetic variance that ends up as residual error in GWAS (Shirali et al. 2018). In the present study, we aimed to apply haplotype GWAS to the Pee Dee germplasm enhancement program.

Specifically, the Pee Dee program has a long history of fiber quality improvement, especially fiber strength, as well as an emphasis on improving other fiber traits and yield components. Previously reported replicated field trials provide an extensive catalogue of these phenotypic traits across four states for a total of 14 year-location environments (Campbell et al. 2009). Diversity analysis based on SSR markers (Campbell et al. 2009), as well as prior analysis of these field trials (Campbell et al. 2011; Campbell et al. 2012), revealed differences in trait correlations over time, extensive environmental interference with trait expression, and unique combinations of yield and fiber quality in a few founding germplasm lines. We hypothesized that the haplotype-based GWAS analysis would reveal sets of co-segregating SNPs that underlie these key traits, and that we would be able to track these haplotypes across the extant cotton cultivars, especially since multiple Pee Dee breeding lines were used as donors for fiber strength genes in most of the US germplasm (Bowman and Gutierrez 2003). To that end, the objectives of this study were to 1) characterize the genetic architecture of eleven yield-related and fiber quality traits in the Pee Dee germplasm by identifying haplotypes with a negative or positive effect on these traits; 2) identify and discuss genomic regions with QTL for multiple traits; and 3) study linkage disequilibrium and gene flow to anchor these findings relative to the complex history of the Pee Dee program.

Materials and Methods

A set of 80 Pee Dee genotypes was genotyped on the CottonSNP63K Array (Hulse-Kemp et al. 2015). Other improved upland cotton genotypes, 272 from Hinze et al. (2017) and 16 from Billings et al. (2020), were used to impute and phase missing SNP calls with BEAGLE v5.1 (Browning et al. 2018; Browning and Browning 2007) for this set of 80 Pee Dee genotypes. The set of 80 Pee Dee genotypes examined in this study was separated out and filtered post-imputation [minor allele frequency (MAF) > 2.5%; ≥ 1 individual in each homozygous class]. A thinned set of SNPs, with more uniform marker density across the genome, was generated using the "--indep-pairwise" command in plink (Chang et al. 2015). Inter-marker correlation-based haplotype block estimation was performed with the "--blocks" command in plink. Haplotype blocks were considered nonoverlapping sets of genetic variants whose alleles are usually inherited together (Gabriel et al. 2002). Some SNPs were considered alone if no nearby SNPs were highly correlated. In addition, population structure analysis and the calculation of a kinship matrix were performed using fastSTRUCTURE to account for the relatedness between genotypes in this study (see **Population Structure Analysis** in the Supplemental Methods)

Eighty-two Pee Dee genotypes and two to six commercial check cultivars were grown in six locations for either two or three years from 2004 to2006, for a total of 14 environments in South Carolina, North Carolina, Georgia, and Mississippi (Campbell et al. 2009). These included three locations in South Carolina [Florence (FL04, FL05, FL06), Blackville (BL04, BL05, BL06), Hartsville (HV04, HV05)], one location in North Carolina (RM05, RM06), one location in Georgia (TFT05, TFT06), and one location in Mississippi (ST05, ST06). The trial in each location was carried out in an α-lattice incomplete block design and managed according to recommended growing practices for each environment. Fiber analysis was performed with High Volume Instrument and Advanced Fiber Information System at the Cotton Incorporated Fiber Testing Laboratory (Cary, NC, USA).

Using the method Campbell et al. (2009) originally implemented for this dataset, adjusted phenotypic means were calculated for each of the fourteen traits with a custom macro for PROC MIXED in SAS 9.4 (SAS Institute, Cary, NC, USA) with REML. To get an estimate of each genotype’s performance across a wide range of environments, least squares means were calculated for the genotypes with the following model:

$$P = U + G + YL + BLK(YL) + G * YL + E \qquad \textbf{\textit{Eq. 1}}$$

where P is the estimated mean, U is the overall mean, G is the fixed genotype effect, and random effects YL (effect of that year-location), BLK(YL) (incomplete block nested in year-location), G*YL (interaction between genotype and year-location), and residual error, E. Like Campbell et al. (2009), we also calculated least square means at the individual year-location level. Only those phenotypic means estimated from a dataset with a significant *F*-statistic ($p < 0.05$) for the genotype effect were included for GWAS.

Haplotype-based GWAS analysis was performed with the RAINBOW model, as implemented in the function "RGWAS.multisnp" in the R package 'RAINBOWR' (Hamazaki and Iwata 2020). A linear kernel-based association test was employed following the approach of Hamazaki and Iwata (2020), which is estimated local to each haplotype block using the natural and orthogonal interactions (NOIA) method (Vitezica et al. 2017). The NOIA estimates of genetic variance underlying a phenotype have the key advantage of allowing separate inferences for additive, dominance, and epistatic interactions, as well as reduced skew from markers in high LD or out of Hardy-Weinberg equilibrium. Because we were interested in additive genetic effects, we chose the additive portion of genetic variance as partitioned by NOIA. A simplified version of the RAINBOW model is provided here:

$$P = X\beta + u_c + u_i + E \qquad \textbf{\textit{Eq. 2}}$$

Where P is the phenotypic value, Xβ is the vector of fixed effects and model intercepts, including those associated with the fastSTRUCTURE Q matrix, u_c is the vector of random background genetic effects derived from the kinship matrix K, u_i is the vector of random effects associated with the i-th SNP-set estimated by transforming the local genotypic matrix, and E is random error. The haplotype-set test estimates variance components for the model with SNPs as random effects using the eigen decomposition of the local genetic relationship matrix. The likelihood ratio test, where the null model excludes the SNP-set of interest, was used to estimate the *p*-value of a given SNP-set by testing for significant improvement of model fit.

For those datasets with a significant *F*-statistic for the genotype effect, this likelihood ratio test was performed for each haplotype block to test for association with all eleven traits for the least squares means for the 14 year-locations separately and the overall means across all year-locations combined. Haplotype blocks with a *p*-value less than Bonferroni correction (i.e. $p < 0.05$/number of blocks) were designated as significant haplotype blocks. Because of occasional missing data, some SNPs were discarded due to low MAF after removing individuals with missing phenotypic data.

Haplotype-based GWAS analysis identified significantly associated haplotype blocks but did not explain which single variant, or set of variants, in that chromosomal segment conferred a positive or negative effect on the phenotype. To determine which genotypes were associated with the variation for a phenotypic trait, the results from single marker analysis were first examined to see if a significant SNP marker was present in the haplotype region. If present, no further multi-marker analysis was performed in that region. For this single marker analysis, each genotype was grouped into one of three classes (homozygous for the common allele, heterozygous, or homozygous for the minor allele) and an *F*-test for the effect of the marker was performed. If no significant single markers were present in this region, hierarchical clustering was used to group together similar haplotypes. Subsequently, ANOVA was performed to test for association between haplotype clusters and trait variation. Lastly, if there was no significant effect due to cluster membership, the genotypes were separated by haplotype and a *t*-test

(α=0.05) was performed to identify variants associated with superior or inferior phenotypic values (Li et al. 2020).

The least common version of each significant haplotype was marked as having an increased or decreased effect as compared to the most common variants at that locus by examining the results of the pairwise *t*-test. These results were visualized using boxplots for the phenotype, separated by the appropriate grouping method, and examining the direction of each group's effect. The percent residual variance explained for the full model including all the significant haplotypes was calculated using an R^2 measure based on the likelihood ratio test (Nagelkerke 1991), where the null model was the mixed linear model with no markers. The calculation was done with the "r.square.LR" function in the R package 'MuMin' (Barton 2020).

Hierarchical clustering was performed with the "hclust" and "cuttree" functions in R to group together similar haplotype variants (maximum number of groups = 3). Separation into unique haplotypes was performed by concatenating all of the SNPs together and grouping by unique haplotypes.

Results and Discussion

Analysis began with a total of 14 year-locations of raw data for which eleven phenotypic traits were collected, including five yield-related components, four HVI fiber quality parameters, and two AFIS fiber quality parameters (**Table 3.1**). A simple, single marker analysis was first attempted with just the first two dimensions of PCA as covariates, but high genomic inflation factors (λ>2) associated with long-range LD resulted in very low statistical power after applying the appropriate correction (Yang et al. 2011). As a result, the kinship matrix and fastSTRUCTURE membership probabilities (*k*=4) were added to the final GWAS model, with linked markers clustered into haplotypes and tested one block at a time.

The haplotype blocks varied significantly in length, ranging from a 75Mb haplotype block containing 1,152 SNPs (chromosome A08) to a pair of SNPs that were 30bp apart (A09). Across the 1,751 haplotype blocks discovered, 75 spanned a length <1kb, 228 were in the 1kb-10kb range, 603 10kb-100kb, 687 100kb-1Mb, and 158 > 1Mb (**Figure 3.1**). These haplotype block span size estimates are similar to those described elsewhere (Abdullaev et al. 2017). Sporadic and extensive LD structure was previously observed in genotypes sourced from the Pee Dee breeding program (Billings et al. 2020), and these observations were confirmed on the subset of genotypes studied here. In addition to the 1,751 haplotype blocks containing two or more SNPs, an additional 1,487 SNPs were assigned to their own haplotype block due to absence of adjacent markers in LD. One consequence of the LD structure of this data set is that mapping resolution can either be very fine or very poor, depending on whether or not recombination has occurred historically at a given locus. Strong selection over the course of the breeding program may have resulted in reduced LD in genomic regions underlying key traits, perhaps having the opposite effect on mapping resolution.

A linear kernel-based association analysis revealed 66 significant haplotype block-trait associations. Among these haplotype blocks, 15 trait associations were found in the ALL dataset and 51 for traits measured in one of the fourteen individual environments (**Table 3.2** and **Figure 3.2**). The greatest number of significant haplotype blocks were detected in the ALL and ST05 datasets (15 and 14, respectively). No associations were found either year in Rocky Mount (RM05, RM06), in two of three years in Blackville (BL05, BL06), or in one year at Tifton (TFT05). Because duplicate haplotype blocks were discovered in different environments, a total of 33 unique haplotype blocks were found to be significantly associated with at least one trait (**Table 3.3**).

The QTL hits from single marker analysis were compared to the haplotype GWAS to look for common genome regions detected in both. Thirty-seven of the 66 haplotype blocks were shared between the two methods. The remaining 29 haplotype blocks were identified with haplotype GWAS although no single SNPs in each block was cross-validated. Example Manhattan plots of haplotype blocks passing and failing single marker analysis cross-validation are given in **Figure 3.3**.

There are a few explanations for why an entire block may be significant, but the individual SNPs are not. The haplotype block analysis may be excluding false positive QTL (and true QTL) suggested by single marker analysis because of the pooling of adjacent SNPs in the local genomic relationship matrix (Hamazaki and Iwata 2020). Likewise, the score function applied in RAINBOW with the NOIA kernel is affected by the background frequency of a variant, so undue significance is not given to a single rare variant in a haplotype block.

These significant haplotypes with an overlapping signal from single marker analysis may indicate QTL of large effect, where a single SNP (or adjacent SNPs) is suitable for capturing the underlying genetic variation at a locus contributing to the phenotypic value. For the remaining 29 haplotype blocks awaiting cross-validation, other genetic patterns were explored that may explain the observed GWAS signal. Where a single SNP was not adequate, 19 significant haplotype blocks were classified with clustering analysis and eleven were separated by unique haplotypes. A list of findings from each of these steps is in **Supplemental Table 3.2**, and example boxplots for these three categories is given in **Figure 3.4**.

The percent residual variance explained (PVE) was calculated for each set of haplotype blocks for a trait in an environment (**Table 3.4**). The PVE ranged from a low of 5% for bolls m^{-2} in ST06 (for one haplotype block) to a high of 60.7% for upper half mean length in ST06. For most traits, the PVE was around 25%, indicating that the error variance in the whole model was reduced once adding the effect of the significant haplotype blocks. For upper half mean length, the discovery of many high effect, environment-specific QTL contributed to large PVE in most cases. The single highest effect QTL (PVE = 56.1%) was in a haplotype block associated with an increase in upper half mean length in ST06.

In total, five significant haplotypes for yield components, two for lint yield, and 26 for fiber quality traits were discovered. These 33 associations were scattered across 26 independently segregating genomic regions.

Haplotypes Only Associated with Yield and Yield Components

Of the seven haplotype blocks associated with lint yield or yield components, two were not located to the same haplotype blocks for any fiber quality traits. One novel QTL was discovered in the ST06 data for bolls m^{-2} on D11 (44.26-45.40 Mb). Cluster analysis revealed a group comprising 26% of individuals in the study with significantly lower bolls m^{-2} than the most common cluster, made up of 60% of individuals. The other QTL, confirmed by single marker analysis with *i52326Gb* (chromosome A12, 106.45 Mb), was associated with a nominal increase in the seed index in the ALL data for the three heterozygotes and one homozygote with the *T* allele. However, two QTL for seed index on either side of this marker were previously reported, suggesting that this is likely a genuine association signal [*qSI-Pop1-A12-1* (Zhang et al. 2016) and *qSI-Chr12-1.XZ.E2-RIL* (Shang et al. 2016a))].

Haplotypes Only Associated with Fiber Quality Traits

Twenty of the 26 fiber quality QTL were not located to the same haplotype blocks for lint yield or yield components. These QTL were distributed across eight chromosomes in thirteen genomic regions. On chromosome A04, two haplotype blocks composed of single SNPs were detected. A QTL for upper half mean length (87.53 Mb) was significant in BL04, FL05, ST05, and ALL. A nearby marker significant for strength (87.70 Mb) was detected in FL04. Both of these QTL have been previously reported independently [*qFL-A5-1.env1* (Shen et al. 2006) and q*FS-chr04-1.15ALE* (Liu et al. 2018)]. The beneficial alleles for both QTL are in repulsion except in the case of a few unique recombinants, in line with a recent report on related material that shows a typically negative correlation between these two traits in segregating populations (Campbell 2020). While Sealand-542 only has the beneficial fiber length allele and PD 2164 only has the strength allele, Hybrid 330-278 contains both beneficial alleles. At the time when Hybrid 330-278 was released, Culp and Harrell (1980) noted that Hybrid 330-278 was one of the first products from their breeding program with combined strength and length, with the length and strength both originating from a complex cross that included Sealand 542 and the parents for PD 2164 (Harrell 1974).

Examination of allele frequencies across all of the improved upland cotton germplasm SNP data present in CottonGen revealed a frequency of 92% having neither beneficial allele; <1% (3) only the strength allele; 4% (15) only the length allele; and the remaining 3% (8) having both beneficial alleles. Interestingly, four of the ten genotypes with both beneficial alleles originated from the Pee Dee program [Hybrid 330-278 and PD 5582 from this study, 'PD-1' and Sealand-7 Yellow Flower from Hinze et al. (2017)], three more are from the Coker breeding program (Calhoun et al. 1997), two had pedigrees that could not be determined ('Dekalb 220' and 'Locket 1'), and the remaining genotype, 'Tidewater-29', is a reselection from one of the founding germplasm lines in the Acala breeding program, which also includes triple hybrid germplasm in its foundation (Zhang et al. 2005). Previous research has suggested that much of the beneficial gain in fiber strength and length can be attributed to these two programs, especially regarding the breaking of the negative linkage between fiber quality and agronomic performance (Culp et al. 1979). Six of the seven genotypes with both beneficial alleles had a Sea Island (*Gossypium barbadense* L.) ancestor somewhere in their pedigrees, indicating this haplotype may have been introgressed from *G. barbadense* L.

A single QTL for micronaire was discovered on chromosome A05 (109.45-109.46 Mb) in BL04. One third of the genotypes in the study belonged to the beneficial cluster at this haplotype block, which had significantly lower micronaire than either of the other two clusters. On the other hand, a QTL for micronaire with a deleterious effect (higher micronaire) minor allele (5% frequency) was detected in FL06 on chromosome D10 (55.52 Mb). Chromosome D05 contained two unlinked QTLs for fiber fineness and upper half mean length. The QTL for fineness (31.89-31.91 Mb) was significant only in BL04; both the *T* allele homozygotes and heterozygotes had higher fiber fineness, indicating a potential dominance effect at this locus for an undesirable change in fineness (this is **A1** in **Figure 3.4**). The haplotype block for upper half mean length on chromosome D05 was detected in ST06 and ALL, with the *TTGAC-GAAACGCCA* present in four of the top eight longest fiber lines. The haplotype blocks are shown as all of the SNPs in that region joined together, with dashes '-' representing that the individuals was heterozygous for that SNP.

Chromosome D06 harbored multiple linked and unlinked QTL for upper half mean length. A small cluster (~14% of individuals) for haplotype block 3004 (22.56-24.28 Mb) was associated with significantly decreased fiber length in ALL. Another nearby association detected in ALL was haplotype block 3005, which spanned 73 SNPs (24.31-44.42 Mb) including a previously reported QTL region [*qFL-D6-1.env2* (Shen et al. 2006)]. The homozygous *T* allele group was associated with increased fiber length. Seven Mb away on the other side of the centromere, an additional four linked haplotype blocks (3008, 3010, 3011, and 3012) were associated with upper half mean length. All four blocks (51.37-57.72 Mb) were significant in ALL and at least four different environments, indicating the ability to discover this QTL in a wide range of environments. Another segment (62.17 Mb) was significant for upper half mean length only in ST06, and lastly one more near the end of the chromosome in ALL and both years at ST (05, 06). Each of these haplotypes exhibited a similar pattern where the minor haplotype or SNP variant was associated with longer fibers in Sealand-542 and Hybrid 330-278. In four of the five blocks, PD 3246 also carried the beneficial haplotype block. PD 4461Q and PD 8619 also had the two flanking beneficial haplotypes, while PD 4381 only had one. However, none of these three genotypes had nearly as long of fibers as the superior lines, possibly because Sealand-542 and Hybrid 330-278 also contain additional beneficial fiber length alleles located in other regions of the genome. Analysis for block 3010 showed that the individuals that were heterozygous had decreased fiber length.

An additional four regions were detected, containing QTL for upper half mean length distributed across three chromosomes. On A08, two linked haplotypes were discovered in a previously reported region for the FL05 means [*FL3.05CQ* (Zhang et al. 2009)]. The *G* allele in block 729 (122.23 Mb) was present in four of the five longest-fiber lines, with the notable exception of Sealand-542. The same pattern was present for the beneficial haplotype *CAAATAA* for block 731 (122.658-122.807 Mb). PD 9223 also contained the beneficial *G* allele in block 729 but had average fiber length, suggesting the true causal locus may be out of linkage with the SNP marker, *i49570Gh*. Three additional QTL for upper half mean length (block 2813, block 2840/2841, and block 1947) were all previously identified [*qFL-c24.E9* (Wang et al. 2015), *qFL24.2.bb07* (Zhang et al. 2011), and *qFL-C18-3.Ay07* (Jamshed et al. 2016)]. With the exception of block 2813, each QTL exhibited effects in multiple environments.

Co-locating Haplotype Blocks for Yield and Fiber Quality

There were an additional four genomic regions with either overlapping or adjacent significant haplotype blocks for both yield and fiber quality traits. In three of the four genomic regions, the high yielding variants were rarely (if ever) found in the same individuals as the high quality variants. This highlights the difficulty overcoming the negative relationship between yield and fiber quality which results from the genetic linkage of these two traits, typically in repulsion phase (Culp et al. 1979; Meredith and Bridge 1971; Smith and Coyle 1997).

On chromosome D06, two haplotypes (*GGTTAGAAATATATACAAGCTGC* and *GATCAGAAATATATACAGGCTGC*) composed of a block of 23 SNPs (44.58-48.77 Mb) were associated with lower gin turnout in ALL, stronger fibers in FL04, and longer upper half mean length in BL04, ST05, ST06, and ALL. The individuals with this haplotype included Hybrid 330-278, PD 3246, PD4381 (the only genotype with the second haplotype), PD 4461Q, and PD 8619. As with the strength and length QTL on A04, pedigree analysis revealed that 93% (19 individuals) of all the improved upland cotton genotypes with resolvable pedigrees carrying either beneficial haplotype had a Sea Island parent somewhere in their pedigree, suggesting a potential origin for this high fiber quality allele at the expense of yield. The gin turnout QTL was previously identified in an introgression experiment with cotton landraces [*qLP-Pop1-D6-1* (Zhang et al. 2016). The upper half mean length QTL was previously reported in a recombinant inbred line population, although limited marker density resulted in a much larger window than found here [*qFL-D6-1.env2* - (Shen et al. 2006)].

On chromosome D07, the predominant SNPs in two unlinked haplotype blocks for lint yield in HV04 (3.29 Mb) and fiber strength in HV05 (3.77-3.83 Mb) were associated with decreased performance. Tan et al. (2014) also found a QTL for fiber strength on D07 (*qFS16.1.2008*) only in one of four environments, suggesting a significant genotype x environment interaction effect at this locus. All four of the lowest yielding genotypes had the *A* allele at the *i27357Gh* marker, and similarly most of the low strength genotypes had the *A* allele at the *i01410Gh* marker. No genotypes contained both the negative strength allele and the negative yield allele, while six (~12%) contained the deleterious allele for yield and more than 25% contained the low strength allele. Approximately 3% of the genotypes in the extant improved upland germplasm contain both negative alleles at this locus, indicating that the negative variants at this locus may have been selected against in the cotton breeding gene pool.

There is a long established positive relationship between micronaire and yield, although the strong environmental impact on both micronaire and yield complicate the stability of this relationship (Elms et al. 2006; Clement et al. 2012). Therefore, it was no surprise that we identified two adjacent haplotype blocks for micronaire and lint yield on chromosome A13. The micronaire QTL (91.61-92.05 Mb) was discovered only in TFT06, with a small genotype cluster (~7% frequency) having significantly higher micronaire than the other genotypes in that environment. The micronaire QTL was previously reported by Tan et al. (2018), who found a QTL (*qFM13.2.2016CQ*) cluster for fiber strength, elongation, and micronaire in this region. The significant haplotype block for lint yield was ~130 kb away (92.18-93.76 Mb), with those individuals homozygous for the *A* allele at the *i13404Gh* marker having significantly lower yield. About half of the genotypes contained the favorable combination of carrying neither the high micronaire haplotype nor the low yield allele, 32 had only the low yield allele, four clustered only with the high micronaire group, and two genotypes had the deleterious combination of the high micronaire and low yield variants.

Associations were also identified on the proximal end of D13 corresponding to two haplotype blocks. One block included only a single SNP marker, *i20441Gh* (1.18 Mb), while the other block included 10 SNPs in strong LD (1.26-1.46 Mb). Associations with both of these blocks were identified for seed index in ST05. In the second block, a single SNP, *i152288Gb* (1.26 Mb), was significant in single marker analysis. The beneficial SNP alleles for seed index were in perfect LD in this population, with 13 genotypes (Earlistaple-7, FJA, FTA, Hybrid 330-278, PD 111, PD 2164, PD 2165, PD 3246, PD 4381, PD 5377, PD 5472, PD 7496, and PD 9363) homozygous for both beneficial alleles for seed index. In the same haplotype block as *i152288Gb*, a single marker *i12997Gh* (1.43 Mb) was significant for both fiber strength in FL04 and upper half mean length in ST05. The *G* allele at this SNP marker was associated with longer and stronger fibers in eight genotypes, a subset of those with the beneficial seed index SNPs (the same as above excluding PD 111, PD 4381, PD 5472, PD 7496, and PD 9363). The QTL for fiber strength and upper half mean length were previously reported [*qFS-Chr18-1.E1.XZV-BC* (Tan et al. 2018) and *qFL18.1.2016HN* (Shang et al. 2016b)], although the signal for seed index on this end of D13 was not previously reported.

Sealand-542 and PD 259 were heterozygous for both markers. PD 9363 carried the positive haplotype for seed index but was heterozygous at the strength/length marker. PD 5529, PD 6992, and PD 785 were heterozygous for the seed index allele and homozygous for the non-beneficial strength/length marker. Due to the unusual LD pattern around three critical traits, we performed further analysis on this segment on chromosome D13. We surveyed the improved upland cotton germplasm to determine the prevalence of these haplotypes. Both the alleles for higher seed index and longer, stronger fiber were detected in 10% of the genotypes, neither allele was present in 76%, and 4% were heterozygous at one of both loci. Only 2% of the genotypes had only the longer/stronger fiber haplotype, and the remaining 8% had only the markers beneficial for seed index. Examination of the available SNP data did not reveal any obvious recombination events or germplasm introduction responsible for this combination of beneficial variants. Further dissection of this trait locus would require denser genotyping on more individuals in their pedigrees.

Conclusions

In this study, significant haplotypes were identified within the Pee Dee germplasm enhancement program associated with variation for four yield components and four fiber quality parameters. A total of 67 significant haplotype associations were found for eight traits in ten individual environments and the mean combined across all environments, establishing 33 QTL. Nearly half of these associations (16) were not previously reported. Most haplotypes associated with yield components and/or fiber quality were not detected consistently across the 14 environments evaluated in this study indicating the importance of genotype x environment interaction for these QTL. In most environments, >50% of phenotypic variance was left unexplained by our QTL model. The <50% that was explained by the QTL was dominated by a small number of major QTL, underscoring the difficulty in detecting low effect variants in the presence of high effect variants. A crucial series of fiber length, strength, and gin turnout QTL were found on chromosome D06. Many of the genome-wide signals were driven by the presence of significantly lower (or higher) phenotypes for a small number of genotypes, highlighting the power of haplotype association for capturing more rare genetic variants, although the method was still robust for the few haplotypes that were in higher frequency (Ionita-Laza et al. 2013; Hamazaki and Iwata 2020). Phase information and haplotype inference were also used to deduce potential historical introgressions of recombination break points, including coupling and repulsion phases, that may have (in part) broken the negative linkage between fiber strength and yield. Results of this study allow for a better understanding of the QTL landscape underlying key traits in the Pee Dee program's germplasm. Many of these beneficial haplotypes were at low frequency in the improved upland cotton gene pool, indicating the ability to further improve fiber quality by introgressing these variants. Accounting of the genetic basis of key fiber quality traits in this breeding program will help breeders plan future crosses and provide the basis for genomic selection in the Pee Dee germplasm.

References

Abdelraheem, A., H. Elassbli, Y. Zhu, V. Kuraparthy, L. Hinze *et al.*, 2020 A genome-wide association study uncovers consistent quantitative trait loci for resistance to Verticillium wilt and Fusarium wilt race 4 in the US Upland cotton. *Theoretical and Applied Genetics* **133** (2):563-577.

Abdullaev, A. A., I. B. Salakhutdinov, S. S. Egamberdiev, E. E. Khurshut, S. M. Rizaeva *et al.*, 2017 Genetic diversity, linkage disequilibrium, and association mapping analyses of *Gossypium barbadense* L. germplasm. *PloS One* **12** (11):e0188125.

Astle, W., and D. J. Balding, 2009 Population Structure and Cryptic Relatedness in Genetic Association Studies. *Statistical Science* **24** (4):451-471.

Barton, K., 2020 R package 'MuMIn': Multi-Model Inference (version 1.43.17).

Bednarz, C. W., W. D. Shurley, W. S. Anthony, and R. L. Nichols, 2005 Yield, Quality, and Profitablity of Cotton Produced at Varying Plant Densities. *Agronomy Journal* **97**:235-240.

Billings, G. T., M. A. Jones, S. Rustgi, A. M. Hulse-Kemp, and B. T. Campbell, 2020 Population structure and genetic diversity of the Pee Dee Cotton Germplasm Collection. *(In preparation)*.

Bowman, D. T., 2000 Attributes of Public and Private Cotton Breeding Programs. *Journal of Cotton Science* **4**:130-136.

Bowman, D. T., and O. A. Gutierrez, 2003 Sources of Fiber Strenth in the U.S. Upland Cotton Crop from 1980 to 2000. *Journal of Cotton Science* **7**:164-169.

Browning, B. L., Y. Zhou, and S. R. Browning, 2018 A One-Penny Imputed Genome from Next-Generation Reference Panels. *American Journal of Human Genetics* **103** (3):338-348.

Browning, S. R., and B. L. Browning, 2007 Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics* **81** (5):1084-1097.

Calhoun, D. S., D. T. Bowman, and O. L. May, 1997 Pedigrees of Upland and Pima Cotton Cultivars Released Between 1970 and 1995, edited by M.A.F.E. Station.

Campbell, B. T., and M. A. Jones, 2005 Assessment of genotype × environment interactions for yield and fiber quality in cotton performance trials. *Euphytica* **144** (1-2):69-78.

Campbell, B. T., V. E. Williams, and W. Park, 2009 Using molecular markers and field performance data to characterize the Pee Dee cotton germplasm resources. *Euphytica* **169** (3):285-301.

Campbell, B. T., P. W. Chee, E. Lubbers, D. T. Bowman, W. R. Meredith *et al.*, 2011 Genetic Improvement of the Pee Dee Cotton Germplasm Collection following Seventy Years of Plant Breeding. *Crop Science* **51** (3):955-968.

Campbell, B. T., P. W. Chee, E. Lubbers, D. T. Bowman, W. R. Meredith *et al.*, 2012 Dissecting Genotype × Environment Interactions and Trait Correlations Present in the Pee Dee Cotton Germplasm Collection following Seventy Years of Plant Breeding. *Crop Science* **52** (2):690-699.

Campbell, B. T., 2020 Examining the relationship between agronomic performance and fiber quality in ten cotton breeding populations. *Crop Science*.

Chandnani, R., C. Kim, H. Guo, T. Shehzad, J. G. Wallace *et al.*, 2018 Genetic Analysis of *Gossypium* Fiber Quality Traits in Reciprocal Advanced Backcross Populations. *The Plant Genome* **11** (1).

Chang, C. C., C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell *et al.*, 2015 Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**:7.

Christensen, O. F., P. Madsen, B. Nielsen, T. Ostersen, and G. Su, 2012 Single-step methods for genomic evaluation in pigs. *Animal* **6** (10):1565-1571.

Clement, J. D., G. A. Constable, W. N. Stiller, and S. M. Liu, 2012 Negative associations still exist between yield and fibre quality in cotton breeding programs in Australia and USA. *Field Crops Research* **128**:1-7.

Culp, T. W., D. C. Harrell, and T. Kerr, 1979 Some Genetic Implications in the Transfer of High Fiber Strength Genes to Upland Cotton. *Crop Science* **19** (4):481-484.

Culp, T. W., and D. C. Harrell, 1980 Registration of Extra-Long Staple Cotton Germplasm (Reg. No. GP 150 to GP 154). *Crop Science* **20** (2):291-291.

Du, X., G. Huang, S. He, Z. Yang, G. Sun *et al.*, 2018 Resequencing of 243 diploid cotton accessions based on an updated A genome identifies the genetic basis of key agronomic traits. *Nature Genetics* **50** (6):796-802.

Elms, M. K., C. J. Green, and P. N. Johnson, 2006 Variability of Cotton Yield and Quality. *Communications in Soil Science and Plant Analysis* **32** (3-4):351-368.

Fang, D. D., J. Xiao, P. C. Canci, and R. G. Cantrell, 2010 A new SNP haplotype associated with blue disease resistance gene in cotton (*Gossypium hirsutum* L.). *Theoretical and Applied Genetics* **120** (5):943-953.

Fang, D. D., J. N. Jenkins, D. D. Deng, J. C. McCarty, P. Li *et al.*, 2014 Quantitative trait loci analysis of fiber quality traits using a random-mated recombinant inbred population in Upland cotton (*Gossypium hirsutum* L.). *BMC Genomics* **15**:397.

Fang, L., Q. Wang, Y. Hu, Y. Jia, J. Chen *et al.*, 2017 Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits. *Nature Genetics* **49** (7):1089-1098.

Foulk, J., A., G. R. Gamble, H. Senter, and W. R. Meredith, 2007 Commercial Cotton Variety Spinning Study HVI and AFIS Spinning Relationship. *Proceedings of the 2007 Beltwide Cotton Conferences, New Orleans, Louisiana, January 9-12, 2007*:1808-1814.

Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy *et al.*, 2002 The structure of haplotype blocks in the human genome. *Science* **296** (5576):2225-2229.

Granato, I., and R. Fritsche-Neto, 2018 R package 'snpReady': Preparing Genotypic Datasets in Order to Run Genomic Analysis (version 0.9.6).

Hamazaki, K., and H. Iwata, 2020 RAINBOW: Haplotype-based genome-wide association study using a novel SNP-set method. *PLoS Computational Biology* **16** (2):e1007663.

Harrell, D. C., 1974 ARS-S-30: Breeding Quality Cotton and the Pee Dee Experiment Station Florence S.C., edited by USDA.

Hinze, L. L., A. M. Hulse-Kemp, I. W. Wilson, Q. H. Zhu, D. J. Llewellyn *et al.*, 2017 Diversity analysis of cotton (*Gossypium hirsutum* L.) germplasm using the CottonSNP63K Array. *BMC Plant Biology* **17** (1):37.

Huang, C., C. Shen, T. Wen, B. Gao, Zhu *et al.*, 2018 SSR-based association mapping of fiber quality in upland cotton using an eight-way MAGIC population. *Molecular Genetics and Genomics* **293** (4):793-805.

Hulse-Kemp, A. M., J. Lemm, J. Plieske, H. Ashrafi, R. Buyyarapu *et al.*, 2015 Development of a 63K SNP Array for Cotton and High-Density Mapping of Intraspecific and Interspecific Populations of *Gossypium* spp. *G3 (Bethesda)* **5** (6):1187-1209.

Ionita-Laza, I., S. Lee, V. Makarov, J. D. Buxbaum, and X. Lin, 2013 Sequence kernel association tests for the combined effect of rare and common variants. *American Journal of Human Genetics* **92** (6):841-853.

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

Islam, M. S., G. N. Thyssen, J. N. Jenkins, L. Zeng, C. D. Delhom *et al.*, 2016 A MAGIC population-based genome-wide association study reveals functional association of *GhRBB1_A07* gene with superior fiber quality in cotton. *BMC Genomics* **17** (1):903.

Jamshed, M., F. Jia, J. Gong, K. K. Palanga, Y. Shi *et al.*, 2016 Identification of stable quantitative trait loci (QTLs) for fiber quality traits across multiple environments in *Gossypium hirsutum* recombinant inbred line population. *BMC Genomics* **17**:197.

Jenkins, J. N., J. C. Mccarty, and W. L. Parrott, 1990 Effectiveness of Fruiting Sites in Cotton: Yield. *Crop Science* **30** (2):365-369.

Johnson, J., K. Lanclos, S. MacDonald, and L. Meyer, 2020 The World and United States Cotton Outlook in *Agricultural Outlook Forum 2020*.

Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman *et al.*, 2008 Efficient control of population structure in model organism association mapping. *Genetics* **178** (3):1709-1723.

Khan, N. U., K. B. Marwat, G. Hassan, F. Hatullah, S. Batool *et al.*, 2017 Genetic Variation and Heritablity for Cotton Seed, Fiber and Oil Traits in *Gossypium hirsutum* L. *Pakistan Journal of Agricultural Research* **30** (4).

Kim, B., W. D. Beavis, and J. Leon, 2016 Numericware N: Numerator Relationship Matrix Calculator. *Journal of Heredity* **107** (7):686-690.

Korte, A., and A. Farlow, 2013 The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* **9**:29.

Lewis, H., L. May, and F. Bourland, 2000 Cotton yield components and yield stability, pp. 532-536 in *2000 Proceedings Beltwide Cotton Conferences, San Antonio, USA, 4-8 January, 2000: Volume 1.* National Cotton Council.

Li, C., Y. Dong, T. Zhao, L. Li, C. Li *et al.*, 2016 Genome-Wide SNP Linkage Mapping and QTL Analysis for Fiber Quality and Yield Traits in the Upland Cotton Recombinant Inbred Lines Population. *Frontiers in Plant Science* **7**:1356.

Li, Z., P. Kemppainen, P. Rastas, and J. Merila, 2018 Linkage disequilibrium clustering-based approach for association mapping with tightly linked genomewide data. *Molecular Ecology Resources* **18** (4):809-824.

Li, Z., X. Liu, X. Xu, J. Liu, Z. Sang *et al.*, 2020 Favorable haplotypes and associated genes for flowering time and photoperiod sensitivity identified by comparative selective signature analysis and GWAS in temperate and tropical maize. *The Crop Journal* **8** (2):227-242.

Liu, R., J. Gong, X. Xiao, Z. Zhang, J. Li *et al.*, 2018 GWAS Analysis and QTL Identification of Fiber Quality Traits and Yield Components in Upland Cotton Using Enriched High-Density SNP Markers. *Frontiers in Plant Science* **9**:1067.

Ma, Z., S. He, X. Wang, J. Sun, Y. Zhang *et al.*, 2018 Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield. *Nature Genetics* **50** (6):803-813.

Meredith, W. R., and R. R. Bridge, 1971 Breakup of Linkage Blocks in Cotton, *Gossypium hirsutum* L. *Crop Science* **11** (5):695-698.

Meredith, W. R., and R. Wells, 1989 Potential for Increasing Cotton Yields through Enhanced Partitioning to Reproductive Structures. *Crop Science* **29** (3):636-639.

Morota, G., and D. Gianola, 2014 Kernel-based whole-genome prediction of complex traits: a review. *Frontiers in Genetics* **5**:363.

Nagelkerke, N. J. D., 1991 A note on a general definition of the coefficient of determination. *Biometrika* **78** (3):691-692.

Naoumkina, M., G. N. Thyssen, D. D. Fang, J. N. Jenkins, J. C. McCarty *et al.*, 2019 Genetic and transcriptomic dissection of the fiber length trait from a cotton (*Gossypium hirsutum* L.) MAGIC population. *BMC Genomics* **20** (1):112.

Odong, T. L., J. van Heerwaarden, J. Jansen, T. J. van Hintum, and F. A. van Eeuwijk, 2011 Determination of genetic structure of germplasm collections: are traditional hierarchical clustering methods appropriate for molecular marker data? *Theoretical and Applied Genetics* **123** (2):195-205.

Paterson, A. H., Y. Saranga, M. Menz, C. X. Jiang, and R. J. Wright, 2003 QTL analysis of genotype x environment interactions affecting cotton fiber quality. *Theoretical and Applied Genetics* **106** (3):384-396.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick *et al.*, 2006 Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38** (8):904-909.

Raj, A., M. Stephens, and J. K. Pritchard, 2014 fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* **197** (2):573-589.

Shang, L., A. Abduweli, Y. Wang, J. Hua, and J. Jenkins, 2016a Genetic analysis and QTL mapping of oil content and seed index using two recombinant inbred lines and two backcross populations in Upland cotton. *Plant Breeding* **135** (2):224-231.

Shang, L., Y. Wang, X. Wang, F. Liu, A. Abduweli *et al.*, 2016b Genetic Analysis and QTL Detection on Fiber Traits Using Two Recombinant Inbred Lines and Their Backcross Populations in Upland Cotton. *G3 (Bethesda)* **6** (9):2717-2724.

Shen, X., W. Guo, Q. Lu, X. Zhu, Y. Yuan *et al.*, 2006 Genetic mapping of quantitative trait loci for fiber quality and yield trait by RIL approach in Upland cotton. *Euphytica* **155** (3):371-380.

Shin, J., and C. Lee, 2015 A mixed model reduces spurious genetic associations produced by population stratification in genome-wide association studies. *Genomics* **105** (4):191-196.

Shirali, M., S. A. Knott, R. Pong-Wong, P. Navarro, and C. S. Haley, 2018 Haplotype Heritability Mapping Method Uncovers Missing Heritability of Complex Traits. *Scientific Reports* **8** (1):4982.

Smith, C. W., and G. G. Coyle, 1997 Association of Fiber Quality Parameters and Within-Boll Yield Components in Upland Cotton. *Crop Science* **37** (6):1775-1779.

Su, J., S. Fan, L. Li, H. Wei, C. Wang *et al.*, 2016 Detection of Favorable QTL Alleles and Candidate Genes for Lint Percentage by GWAS in Chinese Upland Cotton. *Frontiers in Plant Science* **7**:1576.

Sun, G., C. Zhu, M. H. Kramer, S. S. Yang, W. Song *et al.*, 2010 Variation explained in mixed-model association mapping. *Heredity* **105** (4):333-340.

Tan, Z., X. Fang, S. Tang, J. Zhang, D. Liu *et al.*, 2014 Genetic map and QTL controlling fiber quality traits in upland cotton (*Gossypium hirsutum* L.). *Euphytica* **203** (3):615-628.

Tan, Z., Z. Zhang, X. Sun, Q. Li, Y. Sun *et al.*, 2018 Genetic Map Construction and Fiber Quality QTL Mapping Using the CottonSNP80K Array in Upland Cotton. *Frontiers in Plant Science* **9**:225.

Tang, B., J. N. Jenkins, C. E. Watson, J. C. J. McCartey, and R. G. Creech, 1996 Evaluation of genetic variances, heritabilities, and correlations for yield and fiber traits among cotton F2 hybrid populations. *Euphytica* **91**:315-322.

Thyssen, G. N., J. N. Jenkins, J. C. McCarty, L. Zeng, B. T. Campbell *et al.*, 2019 Whole genome sequencing of a MAGIC population identified genomic loci and candidate genes for major fiber quality traits in upland cotton (*Gossypium hirsutum* L.). *Theoretical and Applied Genetics* **132** (4):989-999.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. *Journal of Dairy Science* **91** (11):4414-4423.

Velazco, J. G., M. Malosetti, C. H. Hunt, E. S. Mace, D. R. Jordan *et al.*, 2019 Combining pedigree and genomic information to improve prediction quality: an example in sorghum. *Theoretical and Applied Genetics* **132** (7):2055-2067.

Viator, R. P., R. C. Nuti, K. L. Edmisten, and R. Wells, 2005 Predicting cotton boll maturation period using degree days and other climatic factors. *Agronomy Journal* **97** (2):494-499.

Vitezica, Z. G., A. Legarra, M. A. Toro, and L. Varona, 2017 Orthogonal Estimates of Variances for Additive, Dominance, and Epistatic Effects in Populations. *Genetics* **206** (3):1297-1307.

Waksmunski, A. R., L. R. Main, and J. L. Haines, 2020 Segregation, linkage, GWAS, and sequencing, pp. 7-23 in *Genetics and Genomics of Eye Disease*.

Wang, H., C. Huang, H. Guo, X. Li, W. Zhao *et al.*, 2015 QTL Mapping for Fiber and Yield Traits in Upland Cotton under Multiple Environments. *PloS One* **10** (6):e0130742.

Wang, L., P. Jia, R. D. Wolfinger, X. Chen, and Z. Zhao, 2011 Gene set analysis of genome-wide association studies: methodological issues and perspectives. *Genomics* **98** (1):1-8.

Wang, S. B., J. Y. Feng, W. L. Ren, B. Huang, L. Zhou *et al.*, 2016 Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Scientific Reports* **6**:19444.

Wu, M. C., S. Lee, T. Cai, Y. Li, M. Boehnke *et al.*, 2011 Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics* **89** (1):82-93.

Yang, H. C., H. Y. Hsieh, and C. S. Fann, 2008 Kernel-based association test. *Genetics* **179** (2):1057-1068.

Yang, J., M. N. Weedon, S. Purcell, G. Lettre, K. Estrada *et al.*, 2011 Genomic inflation factors under polygenic inheritance. *European Journal of Human Genetics* **19** (7):807-812.

Zhang, J. F., Y. Lu, H. Adragna, and E. Hughs, 2005 Genetic Improvement of New Mexico Acala Cotton Germplasm and Their Genetic Diversity. *Crop Science* **45** (6):2363-2373.

Zhang, K., J. Zhang, J. Ma, S. Tang, D. Liu *et al.*, 2011 Genetic mapping and quantitative trait locus analysis of fiber quality traits using a three-parent composite population in upland cotton (*Gossypium hirsutum* L.). *Molecular Breeding* **29** (2):335-348.

Zhang, S., L. Feng, L. Xing, B. Yang, X. Gao *et al.*, 2016 New QTLs for lint percentage and boll weight mined in introgression lines from two feral landraces into *Gossypium hirsutum* acc TM-1. *Plant Breeding* **135** (1):90-101.

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

Zhang, Z., E. Ersoz, C. Q. Lai, R. J. Todhunter, H. K. Tiwari *et al.*, 2010 Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics* **42** (4):355-360.

Figures and Tables

Table 3.1. A summary of phenotypic data collected in fourteen environments across Mississippi, Georgia, South Carolina, and North Carolina from 2004-2006. Agronomic data include lint percent (GIN), lint yield (LYLD), boll size (BWT), seed index (SI), and bolls per square meter (BM2). Fiber quality traits include micronaire (MIC), upper half mean length (UHML), strength (STR), fineness (FINE), and maturity ratio (MATR). Environments with a non-significant genotype effect on a trait are labeled NS. Environments with data not collected are labeled NA.

Year-Location	GIN	LYLD	BWT	SI	BM2	MIC	UHML	UI	STR	FINE	MATR
ALL	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Blackville 2004	✓	NS	NS	✓	NS	✓	✓	✓	✓	✓	✓
Blackville 2005	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Blackville 2006	✓	✓	✓	✓	✓	✓	✓	✓	✓	NS	✓
Florence 2004	✓	✓	NS	✓	NS	✓	✓	✓	✓	✓	✓
Florence 2005	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Florence 2006	✓	NA	✓	✓	NA	✓	✓	NS	✓	✓	✓
Hartsville 2004	✓	✓	✓	✓	NS	NS	✓	✓	✓	NS	✓
Hartsville 2005	✓	NA	✓	✓	NA	✓	✓	✓	✓	✓	✓
Rocky Mount 2005	✓	✓	✓	✓	NS	✓	✓	NS	✓	✓	NS
Rocky Mount 2006	✓	✓	✓	✓	✓	NS	✓	✓	✓	✓	NS
Stoneville 2005	✓	✓	NS	✓	NS	NS	✓	✓	✓	NS	NS
Stoneville 2006	NS	✓	NS	✓	✓	NS	✓	NS	✓	NS	NS
Tifton 2005	✓	NS	✓	✓	NS	✓	✓	NS	✓	NS	NS
Tifton 2006	✓	✓	NS	✓	✓	✓	✓	✓	✓	✓	✓

Figure 3.1. Chromosome positions of 1,751 haplotype blocks discovered with PLINK and 1,487 single SNPs without any highly linked nearby SNPs.

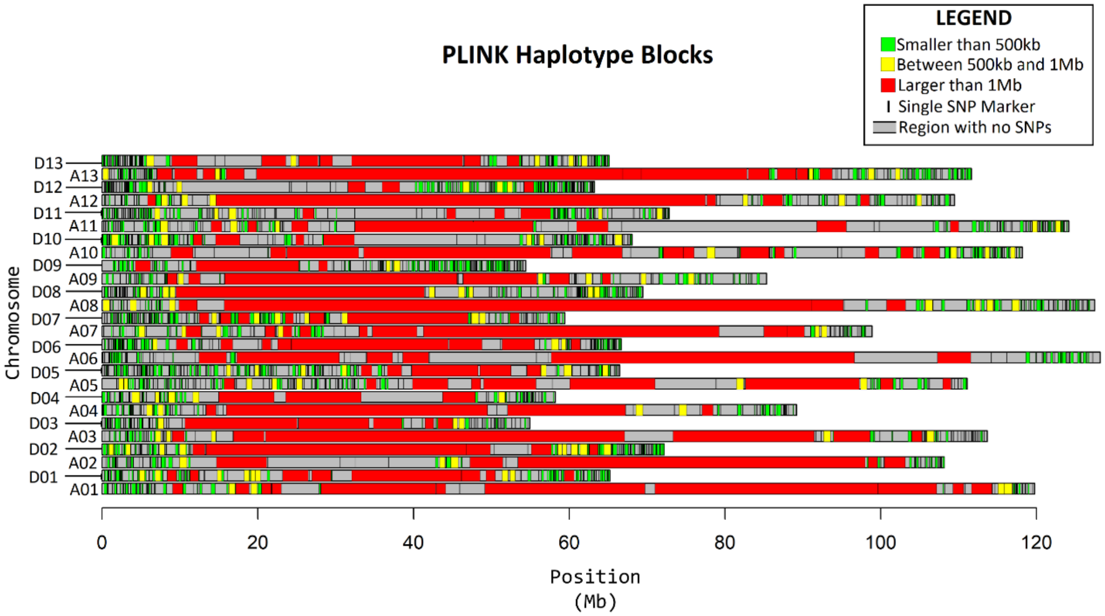


Table 3.2 Number of haplotype blocks associated ($p_{adj\text{-BONF}} < 0.05$) with each trait-environment combination. These include for lint percent (GIN), lint yield (LYLD), boll size (BWT), seed index (SI), and bolls per square meter (BM2). Fiber quality traits include micronaire (MIC), upper half mean length (UHML), strength (STR), fineness (FINE), and maturity ratio (MATR). Sum is the total number of associations found for that row or column, and unique is the number of haplotype blocks located across the genome associated with that trait.

Environment	GIN	LYLD	BWT	SI	BM2	MIC	UHML	UI	STR	FINE	MATR	*Sum*
ALL	1	0	0	1	0	0	13	0	0	0	0	15
Blackville 2004	0	NS	NS	0	NS	1	9	0	0	1	0	11
Blackville 2005	0	0	0	0	0	0	0	0	0	0	0	0
Blackville 2006	0	0	0	0	0	0	0	0	0	NS	0	0
Florence 2004	0	0	NS	0	NS	0	5	0	3	0	0	8
Florence 2005	0	0	0	0	0	0	3	0	0	0	0	3
Florence 2006	0	NA	0	0	NA	1	0	NS	0	0	NS	1
Hartsville 2004	0	1	NS	0	NS	NS	0	NS	0	NS	NS	1
Hartsville 2005	0	NA	NS	0	NA	0	2	0	1	0	0	3
Rocky Mount 2005	0	0	NS	0	NS	0	0	NS	0	0	NS	0
Rocky Mount 2006	0	0	0	0	0	NS	0	0	0	NS	NS	0
Stoneville 2005	0	1	NS	2	NS	NS	11	0	0	NS	NS	14
Stoneville 2006	NS	0	NS	0	1	NS	7	NS	0	NS	NS	8
Tifton 2005	0	NS	0	0	NS	0	0	NS	0	NS	NS	0
Tifton 2006	0	0	NS	0	0	1	1	0	0	0	0	2
Sum	1	2	0	3	1	3	51	0	4	1	0	66
Unique	1	2	0	3	1	3	18	0	4	1	0	

NS indicates trait-environments with a non-significant genotype effect, so they were excluded from GWAS. Cells marked NA did not have phenotypic data available.

Figure 3.2. The genomic locations of haplotypes containing at least one QTL. Traits include lint percent (GIN), lint yield (LYLD), seed index (SI), bolls per square meter (BM2), micronaire (MIC), upper half mean length (UHML), strength (STR), and fineness (FINE).

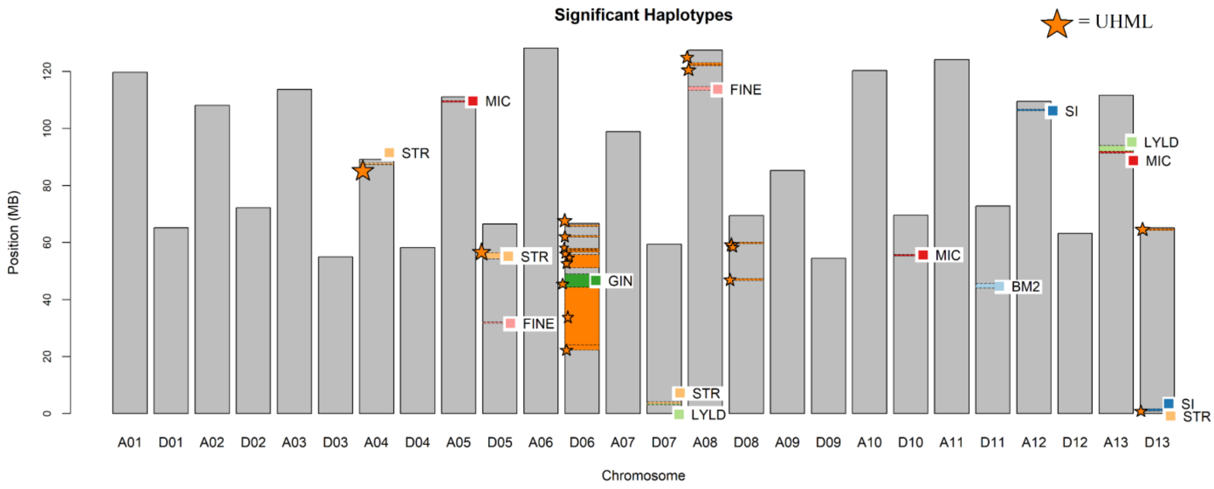


Table 3.3. Summary of 33 haplotype blocks containing a total of 66 QTL discovered in one or more of the fourteen environments.

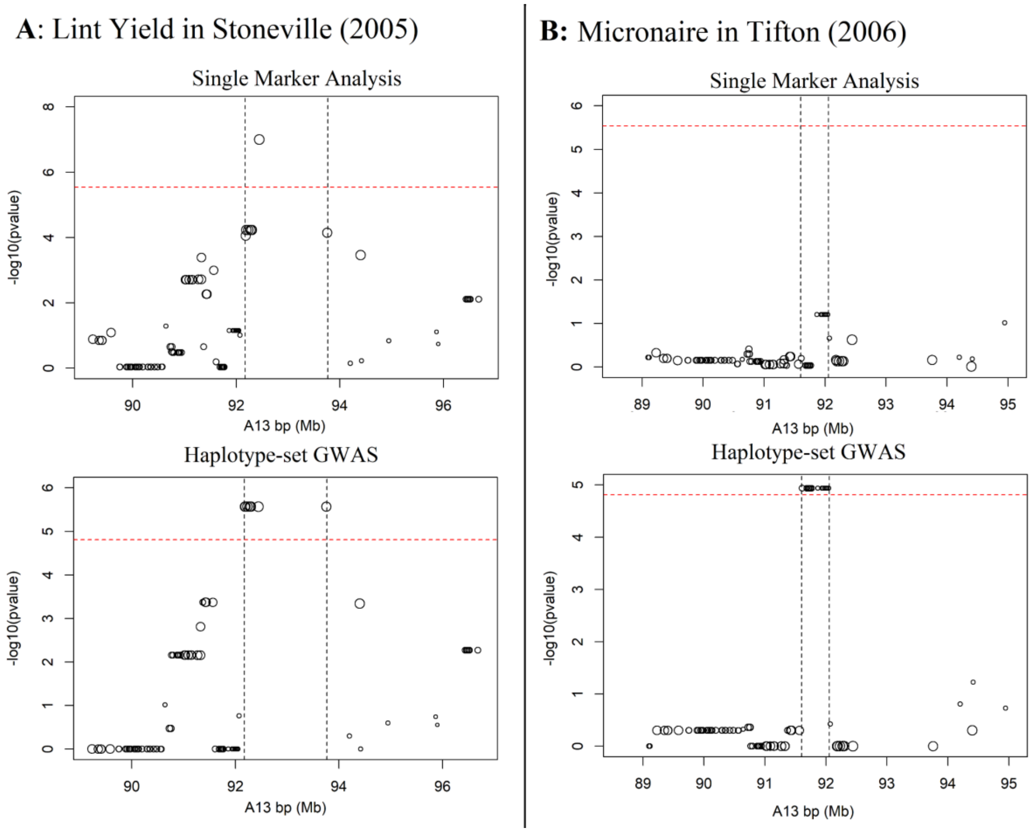
						Trait Associations (in an Environment)	
Block Name	Chr.	Start_BP	End_BP	# SNPs	Size (bp)	Yield Component	Fiber Quality
block334	A04	87,526,177	-	1	-		**UHML** (ALL, BL04, FL05, ST05)
block337	A04	87,659,548	-	1	-		**STR** (FL04)
block484	A05	109,455,966	109,456,985	2	1,019		**MIC** (BL04)
block2112	D05	31,885,973	31,909,118	2	23,145		**FINE** (BL04)
block2129	D05	54,568,741	56,290,754	15	1,722,013		**UHML** (ALL, ST06)
block3004	D06	22,558,506	24,280,462	8	1,721,956		**UHML** (ALL)
block3005	D06	24,311,943	44,419,100	73	20,107,157		**UHML** (ALL)
block3006	D06	44,581,995	48,762,915	23	4,180,920	**GIN** (ALL)	**UHML** (ALL, BL04, ST05, ST06), **STR** (FL04)
block3008	D06	51,365,756	55,530,869	6	4,165,113		**UHML** (ALL, BL04, ST05, ST06)
block3010	D06	56,811,488	57,049,647	11	238,159		**UHML** (ALL, BL04, FL04, ST05)
block3011	D06	57,698,372	-	1	-		**UHML** (ALL, BL04, FL04, HV05, ST05, ST06)
block3012	D06	57,719,266	-	1	-		**UHML** (ALL, BL04, FL04, HV05, ST05, ST06)
block3023	D06	62,167,861	-	1	-		**UHML** (ST06)
block3053	D06	65,908,522	-	1	-		**UHML** (ALL, ST05, ST06)
block1572	D07	3,290,140	-	1	-	**LYLD** (HV04)	
block1577	D07	3,777,721	3,832,375	2	54,654		**STR** (HV05)
block729	A08	122,232,937	-	1	-		**UHML** (FL04)
block731	A08	122,657,930	122,807,477	7	149,547		**UHML** (FL04)
block2813	D08	46,985,484	47,065,574	2	80,090		**UHML** (TFT06)
block2840	D08	59,823,829	-	1	-		**UHML** (ALL, BL04, FL05, ST05)
block2841	D08	59,823,955	-	1	-		**UHML** (ALL, BL04, FL05, ST05)
block2249	D10	55,515,327	-	1	-		**MIC** (FL06)
block2429	D11	44,262,307	45,395,526	2	1,133,219	**BM2** (ST06)	
block1153	A12	106,449,326	-	1	-	**SI** (ALL)	
block1227	A13	91,611,120	92,046,527	18	435,407		**MIC** (TFT06)
block1229	A13	92,180,691	93,755,726	9	1,575,035	**LYLD** (ST05)	
block1838	D13	1,178,229	-	1	-	**SI** (ST05)	
block1840	D13	1,262,304	1,453,941	10	191,637	**SI** (ST05)	**STR** (FL04), **UHML** (ST05)
block1947	D13	64,511,540	64,590,058	3	78,518		**UHML** (ALL, BL04, ST05)

Table 3.4. Percent residual variance explained by significant haplotype blocks. Traits include those for yield components including for lint percent (GIN), lint yield (LYLD), boll size (BWT), seed index (SI), and bolls per square meter (BM2) and for fiber quality traits include micronaire (MIC), upper half mean length (UHML), strength (STR), fineness (FINE), and maturity ratio (MATR). Phenotypic data is from fourteen environments across Mississippi, Georgia, South Carolina, and North Carolina from 2004-2006. The number of significant haplotype blocks discovered in each test is listed in parentheses.

Environment	GIN	LYLD	G25B	SI	BM2	MIC	UHML	UI	STR	FINE	MATR
ALL	23% (1)	-	-	17.8% (1)	-	-	59.2% (13)	-	-	-	-
Blackville 2004	-	NS	NS	-	NS	21.6% (1)	40.9% (9)	-	-	29.4% (1)	-
Blackville 2005	-	-	-	-	-	-	-	-	-	-	-
Blackville 2006	-	-	-	-	-	-	-	-	-	NS	-
Florence 2004	-	-	NS	-	NS	-	44.9% (5)	-	40% (3)	-	-
Florence 2005	-	-	-	-	-	-	30.7% (3)	-	-	-	-
Florence 2006	-	NA	-	-	NA	20.8% (1)	-	NS	-	-	NS
Hartsville 2004	-	29.1% (1)	NS	-	NS	NS	-	NS	-	NS	NS
Hartsville 2005	-	NA	NS	-	NA	-	25.5% (2)	-	27.9% (1)	-	-
Rocky Mount 2005	-	-	NS	-	NS	-	-	NS	-	-	NS
Rocky Mount 2006	-	-	-	-	-	NS	-	-	-	NS	NS
Stoneville 2005	-	33.6% (1)	NS	19.8% (2)	NS	NS	50.6% (11)	-	-	NS	NS
Stoneville 2006	NS	-	NS	-	5% (1)	NS	60.7% (7)	NS	-	NS	NS
Tifton 2005	-	NS	-	-	NS	-	-	NS	-	NS	NS
Tifton 2006	-	-	NS	-	-	0.9% (1)	23.6% (1)	-	-	-	-

NS indicates trait-environments with a non-significant genotype effect, so they were excluded from GWAS. Cells marked NA did not have phenotypic data available.

Figure 3.3. Manhattan plots for single marker analysis and haplotype-set GWAS. A: A QTL discovered in haplotype-set GWAS and in single marker analysis; **B**: A QTL discovered in haplotype-set GWAS and not present in single marker analysis. Point size is proportional to the MAF of the SNP. The red horizonal line is Bonferroni significance ($p < 0.05/\#$ of tests) and the black dashed lines indicate the edges of a haplotype block (this corresponding information derived from the haplotype-set GWAS is only shown in the single marker analysis plots to make the plots easier to compare).



A: Lint Yield in Stoneville (2005)

B: Micronaire in Tifton (2006)

Figure 3.4. Examples of cross-validation for significant haplotypes discovered in GWAS. Boxplots given for single marker analysis (A1/A2), hierarchical clustering (B), and unique haplotypes (C). ANOVA followed by pairwise t-test were used to test for differences between groups.

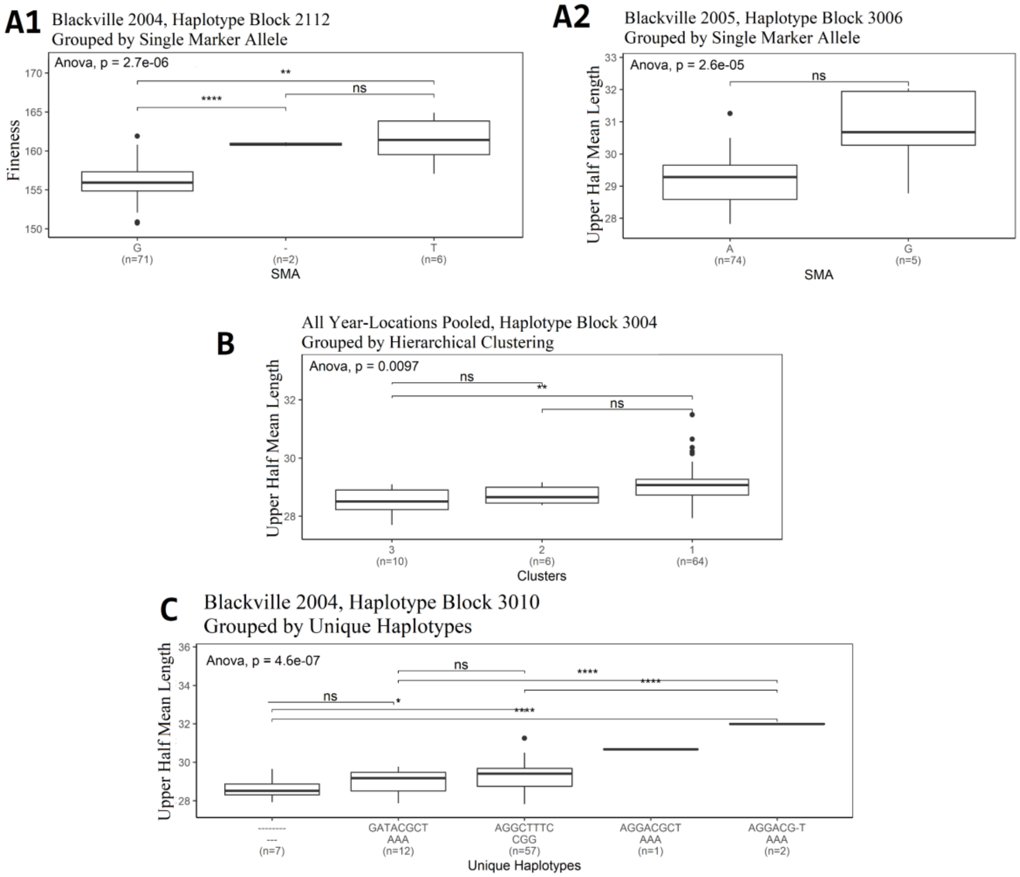
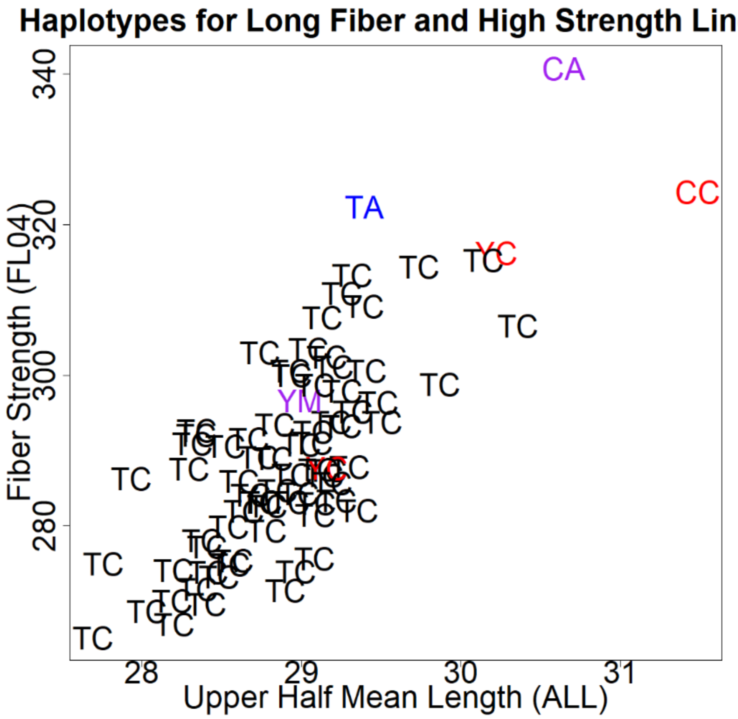


Figure 3.5. The SNP calls for each genotype at the markers associated with fiber strength/length. The observed haplotypes are T and C (T/C); T and A (T/A); C and C (C/C); and C and A (C/A). One individual was heterozygous at both SNPs in the haplotype block (-/-).



Supplemental Methods

Population Structure Analysis

Expanded pedigrees for each of the 81 genotypes included in this study were generated and used to calculate the generalized numerator relationship matrix, 'A', with NumericwareN (Kim et al. 2016). The thinned marker set was used to calculate the additive kinship matrix, 'G', by the first method of VanRaden (2008) with the "G.matrix" function in the R package 'snpReady' (Granato and Fritsche-Neto 2018). The combined 'K' method was used to estimate the individual kinship matrix (Velazco et al. 2019). We used a *w* weighting factor 0.20, corresponding to a kinship matrix that is a weighted average of 20% 'A' and 20% 'G_s', the scaled VanRaden matrix according to average inbreeding in 'A' was estimated by the method in Christensen et al. (2012). 'G_s' can be calculated by solving the following systems of equations:

$$G_s = \beta * G + \alpha \qquad \textbf{\textit{Eq. S3.1}}$$

$$\beta = \frac{Avg(Diag(A)) - Avg(A)}{Avg(Diag(G)) - Avg(G)} \qquad \textbf{\textit{Eq. S3.2}}$$

$$\alpha = Avg(A) - Avg(G) * \beta \qquad \textbf{\textit{Eq. S3.3}}$$

The population substructure matrix, 'Q', was estimated using the fastSTRUCTURE method with default methods for $1 \leq k \leq 10$ (Raj et al. 2014). The optimal number of subpopulations, *k*, was identified by the model complexity that maximized marginal likelihood with the "choosek.py" command.

Supplemental Figures

Supplemental Table 3.1. All non-default settings for programs used in this study.

Task	Program	Command	Flags/Options	Explanation
n/a	plink v1.9	ALL	--autosome-num 26	Sets the chromosome set to 26 chromosomes
			--allow-no-sex	Disables the no-sex warnings
Generating Thinned Data Set		--indep-pairwise	2500 kb	Set the window size to 2.5 Mb
			1	Set step-size at 1 marker, so all adjacent markers are tested
			0.8	Sets the LD threshold (R^2<0.8) for considering SNPs to be independent
Determining Linkage-Based Haplotype Blocks		--blocks	no-pheno-req	Find blocks for all individuals, even with missing phenotypes
		--blocks-max-kb	10000	Find blocks up to 100 Mb in length
		--blocks-min-maf	0.025	Only find blocks with a minimum MAF of 2.5%
		--nonfounders		Include non-founders in the analysis
Phasing and Missing Genotype Imputation	BEAGLE v5.1	gt	= plink.vcf	Reads in the genotype file including 388 improve upland cotton SNP genotypes
		chomr	= i	Selects a single chromosome to run
		out	= imput.chr_i	Sets the output for chromosome i
		window	= 200	Allows BEAGLE to perform imputation on an entire chromosome at once, 200 cM windows (1 cM = 1 Mb)
		ne	= 10000	Effective population size parameter, reduced due to inbreeding
		burnin	= 10	Number of model iterations for determining initial haplotype's [6 is default]
		iterations	= 50	Number of iterations for determining genotype phasing [12 is default]
		phase-states	= 500	Number of model states for genotype phases [280 is default]
		imp-step	= 0.05	Minimum length for small IBS segments [default 0.1 cM; 0.05 corresponds to 50 kb]
		imp-nsteps	= 10	Number of steps used for long IBS segments [default 7]
Determining HMM-Based Phased Haplotypes	HaploBlocker (R pkg v1.5.13)	block_calculation	adaptive_mode = ✓	Repeats model runs to identify haplotypes covering targeted coverage [default 90% coverage/chr]
			consider_multi = ✓	Considers multi-level edges to identify blocks, aid in dealing with phasing inconsistencies [default]
			node_min = 2	Merge even two runs of SNPs into a new block [default 5]
Performing SNP-set GWAS	RAINBOWR (R pkg v0.1.21)	RGWAS.multisnp	ZETA = K	Use the design matrix and additive kinship matrix, estimated as 20% A (pedigrees) and 80% G (vanRaden marker-based kinship)
			structure.matrix = fS	Pass the fastSTRUCTURE Q-matrix (k=6)
			gene.set = plink_blocks	Markers assigned to one haplotype block each, with markers without nearby SNPs in high LD (r2>0.8) assigned to their own block
			min.maf = 0.025	Minimum MAF 2.5% [default 0.02]
			test.method = "LR"	Likelihood-ratio test for estimating p-value for each block
			kernel.method = "linear"	Linear kernel used for estimating local population structure
			test.effect = "additive"	Only test for additive SNP effects

Supplemental Table 3.2. Summary of validation of haploblocks via single marker analysis, haplotype clustering, and separation into unique haplotype blocks. In the second, larger table, the methods used to group plant genotypes for means separation are “SMA” (single marker analysis), “Clusters” (hierarchical clustering on the haplotypes), and “Unique Haps” (separating out in the substituent haplotypes).

	BM2	FINE	GIN	LYLD	MIC	SI	STR	UHML	*SUM*
SMA	0	1	0	2	0	3	3	28	37
Clusters	1	0	1	0	2	0	1	13	18
Unique	0	0	0	0	0	0	0	10	10
** Note: there is 1 haplotype block, MIC (TFT06), that was significant In RAINBOW but did not show up in means separation.									65

									Significant Effect if Grouped By:				
TRAIT	**ENV**	**BLOCK**	**CHR**	**START**	**END**	**#SNP**	**KB**	**Haplo -log10 (P)**	**SMA**	**Clusters**	**Unique Haps**	**Group Method**	**NOTES**
BM2	ST06	block2429	D11	44.262	45.396	2	1133.22	4.82		*TRUE*	*TRUE*	Clusters	Cluster 1 (freq = 26%) has SGFT lower mean
FINE	BL04	block2112	D05	31.886	31.909	2	23.15	4.84	*i25750Gh*	*TRUE*	*TRUE*	SMA	T allele (freq 8%) has SGFT higher mean; - heterozygote (freq 3%) has SGFT higher mean
GIN	ALL	block3006	D06	44.582	48.763	23	4180.92	4.85		*TRUE*	*TRUE*	Clusters	Clust 2 (freq = 8%) has SGFT lower mean
LYLD	ST05	block1229	A13	92.181	93.756	9	1575.04	5.57	*i13404Gh*	*TRUE*	*TRUE*	SMA	A allele (freq 43%) has SGFT lower mean
LYLD	HV04	block1572	D07	3.290	3.290	1	0.00	6.46	*i27357Gh*	*TRUE*	*TRUE*	SMA	A allele (freq 8%) has SGFT lower mean
MIC	TFT06	block1227	A13	91.611	92.047	18	435.41	4.94		*FALSE*	*FALSE*	NONE	- heterozygote (freq 8%) has NS higher mean
MIC	FL06	block2249	D10	55.515	55.515	1	0.00	4.84		*TRUE*	*TRUE*	Clusters	Cluster 2 (freq = 5%) has NS lower mean
MIC	BL04	block484	A05	109.456	109.457	2	1.02	5.31		*TRUE*	*TRUE*	Clusters	Clust 3 (freq = 31%) has SGFT lower mean
SI	ALL	block1153	A12	106.449	106.449	1	0.00	5.24	*i52326Gb*	*TRUE*	*TRUE*	SMA	T allele (singleton) has NS highest mean
SI	ST05	block1838	D13	1.178	1.178	1	0.00	4.88	*i20441Gh*	*TRUE*	*TRUE*	SMA	A allele (freq 16%) has SGFT higher mean
SI	ST05	block1840	D13	1.262	1.454	10	191.64	4.90	*i52288Gb*	*TRUE*	*TRUE*	SMA	A allele (freq 16%) has SGFT higher mean
STR	HV05	block1577	D07	3.778	3.832	2	54.65	4.92	*i01410Gh*	*TRUE*	*TRUE*	SMA	A allele (freq 26%) has SGFT lower mean
STR	FL04	block1840	D13	1.262	1.454	10	191.64	5.27	*i12997Gh*	*TRUE*	*TRUE*	SMA	G allele (freq 9%) has SGFT higher mean
STR	FL04	block3006	D06	44.582	48.763	23	4180.92	5.63		*TRUE*	*TRUE*	Clusters	Cluster 2 (freq = 8%) has SGFT higher mean
STR	FL04	block337	A04	87.660	87.660	1	0.00	5.21	*i49147Gh*	*TRUE*	*TRUE*	SMA	A allele (freq 3%) has NS higher mean
UHML	ST05	block1840	D13	1.262	1.454	10	191.64	5.15	*i12997Gh*	*TRUE*	*TRUE*	SMA	G allele (freq 16%) has SGFT higher mean
UHML	ALL	block1947	D13	64.512	64.590	3	78.52	4.99		*FALSE*	*TRUE*	Unique Haps	T-G haplotype (freq = 3%) NS higher mean
UHML	BL04	block1947	D13	64.512	64.590	3	78.52	4.85		*FALSE*	*TRUE*	Unique Haps	T-G haplotype (freq = 3%) SGFT higher mean
UHML	ST05	block1947	D13	64.512	64.590	3	78.52	4.86		*FALSE*	*TRUE*	Unique Haps	T-G (freq 3%) SGFT highest max
UHML	ALL	block2129	D05	54.569	56.291	15	1722.01	5.09		*FALSE*	*TRUE*	Unique Haps	TTGAC-GAAACGCCA (freq = 5%) SGFT higher mean
UHML	ST06	block2129	D05	54.569	56.291	15	1722.01	7.53		*FALSE*	*TRUE*	Unique Haps	TTGAC-GAAACGCCA (freq = 5%) SGFT higher mean
UHML	TFT06	block2813	D08	46.985	47.066	2	80.09	5.57	*i18770Gh*	*FALSE*	*TRUE*	SMA	A allele (freq 38%) has SGFT higher mean
UHML	ALL	block2840	D08	59.824	59.824	1	0.00	6.55	*i04474Gh*	*TRUE*	*TRUE*	SMA	T allele (freq 3%) has NS highest mean
UHML	BL04	block2840	D08	59.824	59.824	1	0.00	5.10	*i04474Gh*	*TRUE*	*TRUE*	SMA	T allele (freq 3%) has SGFT higher mean
UHML	FL05	block2840	D08	59.824	59.824	1	0.00	4.92		*TRUE*	*TRUE*	Clusters	Cluster 2 (freq = 3%) has NS higher mean
UHML	ST05	block2840	D08	59.824	59.824	1	0.00	5.57		*TRUE*	*TRUE*	Clusters	Cluster 2 (freq = 3%) has SGFT higher mean

UHML	ALL	block2841	D08	59.824	59.824	1	0.00	6.55	*i04475Gh*	*TRUE*	*TRUE*	SMA	C allele (freq 3%) has NS highest mean
UHML	BL04	block2841	D08	59.824	59.824	1	0.00	5.10	*i04475Gh*	*TRUE*	*TRUE*	SMA	C allele (freq 3%) has SGFT higher mean
UHML	FL05	block2841	D08	59.824	59.824	1	0.00	4.92		*TRUE*	*TRUE*	Clusters	Cluster 2 (freq = 3%) has NS higher mean
UHML	ST05	block2841	D08	59.824	59.824	1	0.00	5.57		*TRUE*	*TRUE*	Clusters	Cluster 2 (freq = 3%) has SGFT higher mean
UHML	ALL	block3004	D06	22.559	24.280	8	1721.96	5.57		*TRUE*	*TRUE*	Clusters	Clust 3 (freq 14%) has SGFT lower mean
UHML	ALL	block3005	D06	24.312	44.419	73	20107.16	5.25	*i48830Gh*	*TRUE*	*TRUE*	SMA	T allele (freq 6%) has NS highest mean
UHML	ALL	block3006	D06	44.582	48.763	23	4180.92	6.12	*i48875Gh*	*TRUE*	*TRUE*	SMA	G allele (freq 6%) has NS highest mean
UHML	BL04	block3006	D06	44.582	48.763	23	4180.92	5.67	*i48875Gh*	*TRUE*	*TRUE*	SMA	G allele (freq 6%) has NS highest mean
UHML	ST05	block3006	D06	44.582	48.763	23	4180.92	5.53		*TRUE*	*TRUE*	Clusters	Cluster 2 (freq = 8%) has SGFT higher mean
UHML	ST06	block3006	D06	44.582	48.763	23	4180.92	5.60	*i48875Gh*	*TRUE*	*TRUE*	SMA	G allele (freq 6%) has SGFT higher mean
UHML	ALL	block3008	D06	51.366	55.531	6	4165.11	5.11	*i51081Gb*	*TRUE*	*TRUE*	SMA	G allele (freq 6%) has NS highest mean
UHML	BL04	block3008	D06	51.366	55.531	6	4165.11	5.15	*i51081Gb*	*TRUE*	*TRUE*	SMA	G allele (freq 6%) has NS highest mean
UHML	ST05	block3008	D06	51.366	55.531	6	4165.11	5.20		*TRUE*	*TRUE*	Clusters	Cluster 2 (freq = 8%) has SGFT higher mean
UHML	ST06	block3008	D06	51.366	55.531	6	4165.11	5.15	*i51081Gb*	*TRUE*	*TRUE*	SMA	G allele (freq 6%) has SGFT higher mean
UHML	ALL	block3010	D06	56.811	57.050	11	238.16	5.61		*FALSE*	*TRUE*	Unique Haps	AGGACG-TAAA (freq = 3%) NS higher mean; ------------ (freq = 7%) NS lower mean
UHML	BL04	block3010	D06	56.811	57.050	11	238.16	5.15		*FALSE*	*TRUE*	Unique Haps	AGGACG-TAAA (freq = 3%) SGFT higher mean; ------------ (freq = 7%) NS lower mean
UHML	FL04	block3010	D06	56.811	57.050	11	238.16	5.00		*FALSE*	*TRUE*	Unique Haps	AGGACG-TAA (freq = 3%) SGFT highest mean
UHML	ST05	block3010	D06	56.811	57.050	11	238.16	6.55		*FALSE*	*TRUE*	Unique Haps	AGGACG-TAA (freq = 3%) SGFT highest mean
UHML	ALL	block3011	D06	57.698	57.698	1	0.00	6.75	*i11222Gh*	*TRUE*	*TRUE*	SMA	A allele (freq 4%) has SGFT highest mean
UHML	BL04	block3011	D06	57.698	57.698	1	0.00	6.32	*i11222Gh*	*TRUE*	*TRUE*	SMA	A allele (freq 4%) has SGFT highest mean
UHML	FL04	block3011	D06	57.698	57.698	1	0.00	5.23	*i11222Gh*	*TRUE*	*TRUE*	SMA	A allele (freq 4%) has SGFT higher mean
UHML	HV05	block3011	D06	57.698	57.698	1	0.00	5.13		*TRUE*	*TRUE*	Clusters	Cluster 2 (freq = 4%) has NS higher mean
UHML	ST05	block3011	D06	57.698	57.698	1	0.00	7.60	*i11222Gh*	*TRUE*	*TRUE*	SMA	A allele (freq 4%) has SGFT higher mean
UHML	ST06	block3011	D06	57.698	57.698	1	0.00	5.56	*i11222Gh*	*TRUE*	*TRUE*	SMA	A allele (freq 4%) has SGFT higher mean
UHML	ALL	block3012	D06	57.719	57.719	1	0.00	6.75	*i19972Gh*	*TRUE*	*TRUE*	SMA	A allele (freq 4%) has SGFT highest mean
UHML	BL04	block3012	D06	57.719	57.719	1	0.00	6.32	*i19972Gh*	*TRUE*	*TRUE*	SMA	A allele (freq 4%) has SGFT highest mean
UHML	FL04	block3012	D06	57.719	57.719	1	0.00	5.23	*i19972Gh*	*TRUE*	*TRUE*	SMA	A allele (freq 4%) has SGFT higher mean
UHML	HV05	block3012	D06	57.719	57.719	1	0.00	5.13		*TRUE*	*TRUE*	Clusters	Cluster 2 (freq = 4%) has NS higher mean
UHML	ST05	block3012	D06	57.719	57.719	1	0.00	7.60	*i19972Gh*	*TRUE*	*TRUE*	SMA	A allele (freq 4%) has SGFT higher mean
UHML	ST06	block3012	D06	57.719	57.719	1	0.00	5.56	*i19972Gh*	*TRUE*	*TRUE*	SMA	A allele (freq 4%) has SGFT higher mean
UHML	ST06	block3023	D06	62.168	62.168	1	0.00	4.97	*i28160Gh*	*TRUE*	*TRUE*	SMA	C allele (freq 6%) has SGFT higher mean)
UHML	ALL	block3053	D06	65.909	65.909	1	0.00	5.03		*TRUE*	*TRUE*	Clusters	Clust 2 (freq = 8%) has SGFT higher mean
UHML	ST05	block3053	D06	65.909	65.909	1	0.00	5.21		*TRUE*	*TRUE*	Clusters	Cluster 2 (freq = 8%) has SGFT higher mean
UHML	ST06	block3053	D06	65.909	65.909	1	0.00	5.14	*i17287Gh*	*TRUE*	*TRUE*	SMA	T allele (freq 8%) has SGFT higher mean
UHML	ALL	block334	A04	87.526	87.526	1	0.00	6.49	*i25348Gh*	*TRUE*	*TRUE*	SMA	C allele (freq 3%) has NS highest mean
UHML	BL04	block334	A04	87.526	87.526	1	0.00	5.40	*i25348Gh*	*TRUE*	*TRUE*	SMA	C allele (freq 3%) has SGFT higher mean
UHML	FL05	block334	A04	87.526	87.526	1	0.00	4.99		*TRUE*	*TRUE*	Clusters	Cluster 3 (freq = 3%) has NS higher mean
UHML	ST05	block334	A04	87.526	87.526	1	0.00	5.66		*TRUE*	*TRUE*	Clusters	Clusters 2 (freq = 4%) and cluster 3 (freq = 3%) has SGFT higher mean
UHML	FL04	block729	A08	122.233	122.233	1	0.00	5.24	*i46570Gh*	*TRUE*	*TRUE*	SMA	G allele (freq 6%) has SGFT higher mean
UHML	FL04	block731	A08	122.658	122.807	7	149.55	5.48		*FALSE*	*TRUE*	Unique Haps	CAAATAA (freq = 5%) SGFT highest mean

CHAPTER FOUR

FINAL CONCLUSIONS AND REMARKS

Cotton breeders in the Pee Dee breeding program have managed to breed germplasm lines and cultivars having improved fiber quality while maintaining an adequate standard of yield. In addition, our work here shows that they were able to accomplish those tasks with an apparently narrows genetic base while maintaining genetic diversity and generating novel allelic combinations.

Examination of genome-wide SNP data revealed genetic diversity across 26 chromosomes, although the level of diversity was variable. Multiple population structure evaluation techniques painted a similar picture, which is that clustering and phylogenetic analysis was able to recover some of the original breeding groups in the program, but within-group variation stay approximately constant level over time. Mutations in genes associated with host-plant resistance to disease and insects as well as genes potentially involved in cotton fiber development helped discriminate genotypes from the Pee Dee program compared to other improved upland cotton from around the world.

Haplotype association analysis helped us understand how the genetic variation within the breeding program correlates with fiber quality and field performance. We found that some rare variants from Sea Island cotton likely conferred longer, stronger fiber alleles at the detriment of yield components. Additionally, we found that the predictive capacity of our genetic model highly depended on the environment in which data was collected, implicating a strong genotype by environment effect on all the studied traits in this population.

The results of our work have helped us answer multiple research questions. We expected that genome-wide genetic markers would reflect the history of the breeding program, which we definitely found. However, unexpectedly, we found levels of genetic diversity on-par with much larger samples of upland cotton, suggesting that the breeding techniques and selection methods favored sustained genetic diversity over narrowing of the gene pool. We also found many QTL associated with improved fiber length and strength, but our ability to detect genomic regions underlying other traits was limited, despite ample variation for those traits. External environmental effects or non-additive genetic effects likely impact the ability to detect a signal with SNPs alone.

Despite the limitations of our work, there are many practical applications for continued improvement of cotton. The haplotypes or significant SNPs reported here can be used directly for introgression breeding by anyone who has the ability to score genotypes in their breeding program. The data presented here could also be used in a genomic selection regime to optimize crosses and predict the population sizes necessary to capture rare recombinants for even higher yielding, better-quality cotton. In addition to the plant breeding applications, we have also presented a model that other biologists can use to study diversity in inbred, pedigreed germplasm collections.

APPENDICES

Appendix A

Appendix for Chapters One, Two, Three and Four

Table A.1. List of Genotypes in Chapter Two and Three. The numbers in the group column correspond to the Pee Dee Breeding Group or W if it from the world improve upland cotton germplasm. “2” were used in diversity analysis, “3” in GWAS.

Genotype	Group	Chapter
AC-235	1	2 & 3
AC-241	1	2 & 3
EARLISTAPLE-7	1	2 & 3
EARLISTAPLE-7 (AHK)	1	2
F	1	2
FJA	1	2 & 3
FTA	1	2 & 3
Hy-330-278	1	2 & 3
Sealand-3 (AHK)	1	2
Sealand-542	1	2 & 3
Sealand-542 (AHK)	1	2
Sealand-7-Yellow-Flower (AHK)	1	2
PD2164 (AHK)	2	2
PD2165-242	2	2 & 3
PD2165-242 (AHK)	2	2
PD2165-618	2	2
PD259	2	2 & 3
PD3246	2	2 & 3
PD3249	2	2 & 3
PD4381	2	2 & 3
PD4461Q	2	2 & 3
PD4548	2	2
PD109	3	2 & 3
PD111	3	2 & 3
PD113	3	2 & 3
PD8619	3	2 & 3
PD9223	3	2 & 3
PD9232	3	2 & 3
PD9241	3	2
PD9363	3	2 & 3
PD9364	3	2 & 3
PD9364 (AHK)	3	2
SC-1	3	2 & 3
PD-1	4	2 & 3
PD-1 (AHK1)	4	2
PD-1 (AHK2)	4	2
PD-2	4	2 & 3
PD-2 (AHK)	4	2
PD-3	4	2 & 3
PD-3 (AHK)	4	2
PD6044	4	2 & 3
PD6132	4	2 & 3
PD6179	4	2 & 3
PD6186	4	2 & 3
PD6208	4	2
PD6520	4	2
PD6992	4	2 & 3
PD875	4	2 & 3
PD695	5	2 & 3
PD7388	5	2 & 3
PD7439	5	2 & 3
PD7458	5	2 & 3
PD7496	5	2 & 3
PD7501	5	2 & 3
PD7586	5	2 & 3
PD7723	5	2 & 3
PD781 (AHK)	5	2
PD785 (AHK)	5	2
PD648	6	2 & 3
PD683	6	2 & 3
PD723	6	2 & 3
PD738	6	2 & 3
PD741	6	2 & 3
PD648	6	2 & 3
PD683	6	2 & 3
PD723	6	2 & 3
PD738	6	2 & 3
PD741	6	2 & 3
PD747	6	2 & 3
PD753	6	2 & 3
PD756	6	2 & 3
PD761	6	2 & 3
PD762	6	2 & 3
PD771	6	2 & 3
PD778	6	2 & 3
PD781	6	2 & 3
PD785	6	2 & 3
PD804	6	2
PD878	6	2 & 3
PD948	6	2 & 3
PD5246	7	2 & 3
PD5256	7	2 & 3
PD5256 (AHK)	7	2 & 3
PD5286	7	2 & 3
PD5358	7	2 & 3
PD5363	7	2 & 3
PD5377	7	2 & 3
PD5380	7	2 & 3
PD5472	7	2
PD5529	7	2
PD5576	7	2
PD5582	7	2
PD-3-14	8	2
PD93001	8	2 & 3
PD93001 (AHK)	8	2
PD93002	8	2 & 3
PD93003	8	2
PD93004	8	2 & 3
PD93007	8	2 & 3
PD93007 (AHK)	8	2 & 3
PD93009	8	2
PD93019	8	2 & 3
PD93021	8	2 & 3
PD93030	8	2 & 3
PD93030 (AHK)	8	2 & 3
PD93034	8	2 & 3
PD93043	8	2 & 3
PD93046	8	2
PD93057	8	2
PD94042	8	2
PD94045	8	2
PD97006	8	2
PD97019	8	2
PD97021	8	2
PD97047	8	2 & 3
PD97072	8	2 & 3
PD97100	8	2 & 3
PD97101	8	2 & 3
08-WZ-51	W	2
320F_PI-529233	W	2
4S-180_PI-529496	W	2
A-618	W	2
A-637-33	W	2
Acala-NEM-X1	W	2
Acala-NEM-X2	W	2
Acala-111-Rogers	W	2
Acala-1517-99	W	2
Acala-1517-New-Mexico	W	2
Acala-5_PI-529169	W	2
Acala-Maxxa	W	2
Acala-Royale	W	2
Acala-Ultima	W	2
AK-DJURA-182	W	2
ALA-70-236	W	2
ALBAR-627	W	2
ALBAR-K-603	W	2
ALEPPO-I_PI-529450	W	2
Allen-333_PI-392289	W	2
All-Tex-Atlas	W	2
Arkansas-10	W	2
ARKOT-8102	W	2
ARKOT-8606	W	2
AUBURN-56_PI-529215	W	2
AUBURN-634-RNR	W	2
B163-AH-P9-029-GIBAND	W	2
Beli-Ivzor	W	2
Big-Boll-Triumph	W	2
BJA-592	W	2
BJA-Glandless-Nectariless	W	2
Blightmaster	W	2
BPA-68_PI-365538	W	2
BRS-269	W	2
BRS-286	W	2
BRS-293	W	2
BRS-335	W	2
BRS-336	W	2
BRS-372	W	2
Bulgaria-P73	W	2
CABD3CABCH-1-89	W	2
CABD3SHP3S-1-90	W	2
CAHUGLBBCS-1-88	W	2
Cambodia-4	W	2
CASCOT-B-2	W	2
CD3HCHULBH-1-88	W	2
CD-408	W	2
CD-410	W	2
Central	W	2
Chaco-510-INTA	W	2
Chaco-520	W	2
Christidis-53D7	W	2
Chureza-87	W	2
Ciano-Cocorium-92	W	2
Cleveland-WR-Wannamakers	W	2
CO27GH-Guazuncho-2-Lacape	W	2
Coker-100-Wilt_PI-528761	W	2
Coker-201_PI-529247	W	2
Coker-312_PI-529278	W	2
Coker-312_VanDeynze	W	2
Coker-315_IW-004	W	2
Coker-315_Wilson	W	2
Cokers-Clevewilt-3	W	2
Columbia_PI-528743	W	2
Cook-912-Pope-Clean-Seed	W	2
Cristina	W	2
Dehkanin	W	2
Dekalb-220_PI-529222	W	2
Del-Cerro_PI-529358	W	2
DELCOT-277_PI-529258	W	2
DELTA-OPAL_IW-124	W	2

Genotype	Group	Chapter
DELTA-OPAL_IW-325	W	2
DELTA-OPAL_IW-344	W	2
Deridder-Red	W	2
DES-56	W	2
DES-716	W	2
Dixie-King_PI-529021	W	2
Dixie-Triumph	W	2
Dixie-Triumph-Wannamakers	W	2
DP-10-1	W	2
DP-12_PI-528768	W	2
DP-14_PI-528970	W	2
DP-16_IW-337	W	2
DP-16_PI-529251	W	2
DP-20	W	2
DP-25_PI-529280	W	2
DP-491	W	2
DP-50_PI-529566	W	2
DP-55	W	2
DP-5690	W	2
DP-6_PI-528969	W	2
DP-66_PI-529565	W	2
DP-80	W	2
DP-826	W	2
DP-90	W	2
DP-90_IW-081	W	2
DP-Smoothleaf	W	2
Dunn-219	W	2
Dunn-325	W	2
Empire-WR-61_PI-529224	W	2
Express-121	W	2
Felistana	W	2
FK-290	W	2
FM-832	W	2
FM-993	W	2
FMT-701	W	2
FMT-709	W	2
GA98028	W	2
Garant	W	2
Georgia-King	W	2
Gregg	W	2
Gringo-Inta	W	2
GUAZUNCHO-2	W	2
H-1220	W	2
Half-and-Half_PI-528964	W	2
Hart	W	2
Hopi-Moencopi	W	2
IAC-17	W	2
IAC-18	W	2
IAC-25-RMD	W	2
IMA-12427	W	2
IMA-1318	W	2
IMA-3869	W	2
IMA-6035	W	2
IMACD-8276	W	2
IM-GH	W	2
IRMA-D-742	W	2
Jiangsu-Mian-3_PI-529478	W	2
LA-887_PI-547084	W	2
Lambright-2020-A	W	2
Lankart-57_PI-528822	W	2
LBBCDBOAKH-1-90	W	2
Li1	W	2
Liao-Mian-7	W	2
Lightning-Express	W	2
Limpopo	W	2
Lisina-11	W	2
Lockett-1_PI-529115	W	2
Lockett-BXL	W	2
Lone-Star_PI-528636	W	2
Lu-Mian-14	W	2
Lu-Mian-14_IW	W	2
Lu-Mian-14_IW-074	W	2
M-188-RNR	W	2
M-240	W	2
MAC7-0238	W	2
Magnolia_PI-529033	W	2
MAR5PD208S-4-90	W	2
McNair-210_PI-529589	W	2
McNair-235_PI-529526	W	2
MD-26-NE	W	2
MD51-NE	W	2
MD-52-NE	W	2
MD-90-NE	W	2
Meade-Clean-Seed	W	2
Mebane_PI-528985	W	2
Namcala	W	2
Namcala_IW-314	W	2
NC-88-95	W	2
New-Boykin_PI-528984	W	2
NM24016	W	2
Northern-Star	W	2
NTA-90-8	W	2
Ogosta	W	2
PAK-4F_PI-529301	W	2
PHY-72-Acala	W	2
PHY-PSC-355	W	2
PM-101	W	2
PM-145	W	2
PM-303_PI-529605	W	2
PM-54	W	2
PM-784	W	2
PM-792	W	2
PMHS-200	W	2
PMHS-26	W	2
Pope	W	2
Pora-Inta	W	2
R1TM1-GH	W	2
Reba-B-50_PI-529325	W	2
Reba-P-279	W	2
Reba-P279_AH-531	W	2
Reba-P-288	W	2
Rex	W	2
Riverina-Paplar	W	2
RN-96527	W	2
RN-96625-1	W	2
Rogers-GL-7	W	2
SA-1441_PI-529495	W	2
SA-2330	W	2
SA-2454	W	2
Sabie	W	2
Saenz-Pena-61	W	2
Satu-65_PI-529308	W	2
SG-1001	W	2
SG-747	W	2
Shan-5245	W	2
Shan-5710	W	2
Sicala-3-2	W	2
Sicala-40	W	2
Sicala-40-FM-966	W	2
Sicala-V-2-FM-989	W	2
Sicot-189	W	2
Sicot-53	W	2
Sicot-70	W	2
Sicot-71	W	2
Sicot-81	W	2
Sicot-F-1	W	2
Sicot-F-1_IW-252	W	2
Sioka-1-4	W	2
Soutland-M1	W	2
ST-213_PI-529229	W	2
ST-256	W	2
ST-2C	W	2
ST-453_PI-601544	W	2
ST-474	W	2
ST-825	W	2
Station-Miller-F	W	2
Storm-King-TPSA-1	W	2
TAM-2562-RKNR	W	2
TAM-90J-57S	W	2
TAM-98D-102	W	2
TAMCOT-CAMD-E	W	2
TAMCOT-Luxor	W	2
TAMCOT-Pyramid	W	2
TAMCOT-SP21	W	2
TAMCOT-SP23	W	2
TAMCOT-SP37_IW-142	W	2
TAMCOT-SP37_PI-529637	W	2
TAMCOT-Sphinx	W	2
TASHKENT-I_PI-529447	W	2
TASHKENT-II_PI-529448	W	2
TASHKENT-III_PI-529449	W	2
Tejas	W	2
Tidewater-29	W	2
TM-1	W	2
Toole	W	2
UK-64	W	2
UK-77	W	2
Victoria_PI-606816	W	2
VIR-5850	W	2
VIR-5913	W	2
VIR-6615	W	2
VIR-7263	W	2
Westburn	W	2
Western-Stormproof	W	2
Wilds-18_PI-528781	W	2
Zhong-Mian-Suo-7	W	2
Zhong-Mian-Suo-8	W	2
Zhong-Mian-Suo-9	W	2

Figure A.1. Polymorphic information content for each single SNP or haplotype block. The PIC was calculated by with the “polysat” package in R.

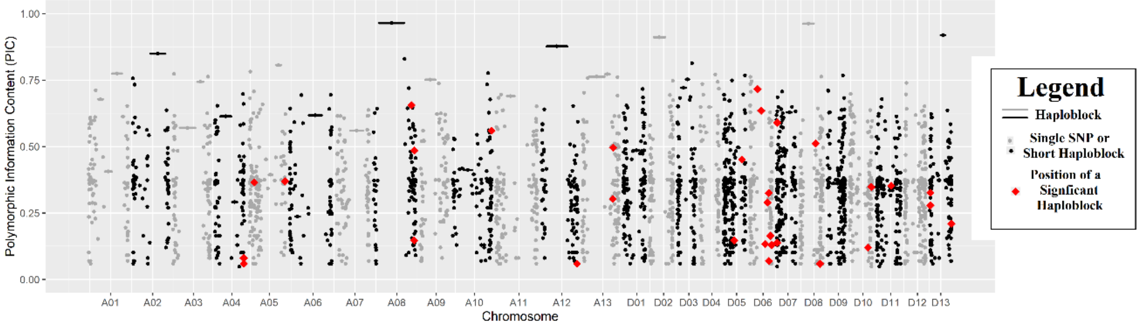


Table A.2. Polymorphic Information Content for the haplotypes in Chapter Three. A higher number indicate better utility for breeding.

Haplotype Block	Polymorphic Information Content
block3004	0.7162
block704	0.6556
block3005	0.6348
block1577	0.5900
block935	0.5588
block2813	0.5114
block1229	0.4956
block731	0.4852
block2129	0.4510
block484	0.3688
block398	0.3648
block2429	0.3526
block2304	0.3475
block1840	0.3268
block3010	0.3254
block1227	0.3031
block3008	0.2896
block1838	0.2787
block1947	0.2095
block3023	0.1638
block2112	0.1469
block729	0.1469
block1572	0.1382
block3006	0.1334
block3053	0.1291
block2249	0.1198
block334	0.0802
block3011	0.0696
block3012	0.0696
block1153	0.0587
block2840	0.0587
block2841	0.0587
block337	0.0587