USING MODERN MOLECULAR AND GENOMIC TOOLS TO EVALUATE POTENTIALLY OVERLOOKED GENETIC VARIATION FOR FIBER QUALITY IN OBSOLETE US IMPROVED COTTON CULTIVARS

A Dissertation

by

MITCHELL J SCHUMANN

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

DOCTOR OF PHILOSOPHY

Chair of Committee,	C. Wayne Smith
Committee Members,	David Stelly
	Lori Hinze
	Alan Dabney
Head of Department,	David Baltensperger

December 2018

Major Subject: Plant Breeding

Copyright 2018 Mitchell John Schumann

ABSTRACT

Improvements in fiber quality phenotyping methods such as High Volume Instrumentation (HVI) and Advanced Fiber Information System (AFIS), have increased breeders ability to detect superior fiber quality; however, this also suggests that potential sources of superior fiber quality prior to the use of HVI and AFIS may have been overlooked. The purpose of this dissertation is to explore gains made in fiber quality according to HVI and AFIS, the stability of these fiber traits, and the feasibility of using genomic prediction to tap into potentially unexploited sources of variation for fiber quality traits.

Genetic gains for HVI and AFIS fiber traits were analyzed using a population of 63 cultivars from the obsolete US improved cotton cultivar collection that represents the past 100 years of breeding efforts in the. All HVI and AFIS traits evaluated made statistically significant gains except for Length by number coefficient of variation, micronaire, and fineness. The statistically significant percent gains per year ranged from 0.267% to 0.025%. Many traits AFIS traits showed gains even though direct selection pressure was likely not applied for these traits, so it is inferred that these gains arose through correlations indicating genetic variation for traits unexploited.

HVI traits and AFIS traits evaluated were highly stable across irrigated and dry environments tested in Corpus Christi and Weslaco, TX in a population consisting of germplasm from obsolete US improved cotton cultivar collection and germplasm with superior fiber quality developed by the Texas A&M Cotton Improvement Laboratory. It was found that selection for these traits in any environment would result in a similar list of genotypes.

Genomic prediction was performed using a population of consisting of germplasm from the obsolete US cotton cultivars collection and germplasm developed for superior fiber quality from the Texas A&M Cotton Improvement Laboratory. Prediction accuracies within the obsolete US cotton cultivars ranged from 0.24-0.56 for HVI and AFIS traits, and variation explained was less than previously reported heritabilities. Prediction accuracy for yarn quality was determined using a selection index created from HVI and AFIS parameters and correlated to yarn work-to-break. Accuracy was determined to be 0.36.

DEDICATION

In dedication to my loving wife Jihye, and our two lovely daughters Camellia and Calliandra.

ACKNOWLEDGEMENTS

I would like to offer my sincerest gratitude to my committee chair Dr. C. Wayne Smith, and my committee members Dr. Lori Hinze, Dr. David Stelly, and Dr. Alan Dabney. All of whom have played an integral role in this research and my professional development. I am grateful for the time they have spent discussing ideas, the professionalism they have displayed, and the example they have lead.

I would like to thank Dawn Deno and all of the graduate students, and student workers in the Cotton Improvement Lab. This project had a large field element, and without their help it would never have been done. I would like also like to thank Dr. Hongbin Zhang and Chantel Scheuring for their help with the lab element of this project. Thank you to Dr. Don Jones and Cotton Incorporated for providing the funding for this research.

Finally, I would like to thank my family. Thank you to my loving wife for standing by me throughout this process, and support and encouragement has made all the difference. Thank you to my two amazing daughters for bringing me so much joy every day and keeping life in perspective for me. Thank you to my parents for teaching me that hard work and perseverance will take you far in life. Thank you to my siblings for helping me to just sit back and laugh at situations sometimes. I love you all very much.

CONTRIBUTORS AND FUNDING SOURCES

This work was supervised by a dissertation committee consisting of Professor C. Wayne Smith (advisor) and Professor David Stelly of the Department of Soil and Crop Science, Professor Alan Dabney of the Department of Statistics, and Dr. Lori Hinze of the United States Department of Agriculture.

The SNP marker array used in this research was completed by Dr. Andrew Hillhouse and Kelli Kochan of the Texas A&M Institute for Genome Sciences and Society. All field related research activities were done with the help of Dr. Steve Hague, Dawn Deno, fellow graduate students, and undergraduate student workers of the Cotton Improvement Lab. All other work conducted for the dissertation was completed by the student independently.

Graduate study was supported by the Diversity Fellowship from Texas A&M University and the Cotton Incorporated Fellowship.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

1. INTRODUCTION

US cotton breeding efforts must keep pace with current demands in cotton fiber quality if US cotton is to remain competitive in the global textile market. These demands come from the improvement of spinning technologies which require longer and stronger fibers for expanded textile portfolios, and faster and more rigorous processing. Current phenotyping advancements, such as High-Volume Instrumentation (HVI) and Advanced Fiber Information System (AFIS), allow breeders to rapidly and objectively quantity fiber quality traits. However, these technologies are recent in the scope of over a hundred years of breeding efforts in US cotton. Many previous methods of analyzing fiber quality were limited to subjectivity, length of time to phenotype, or lack of resources to implement widely into selection platforms. It is possible many sources of fiber quality variation remain unexploited.

As genomic technologies advance, Plant breeders are provided with new tools to facilitate selection of favorable alleles. Genomic prediction is one such tool that utilizes molecular markers to detect genetic variation for the prediction of phenotypic performance. It is different from more traditional QTL analysis in that it is a multi-variate method, and allows for evaluation of multiple marker effects simultaneously (Meuwissen et al., 2001). This technique has application when genotyping is more affordable than phenotyping, and more recently has showed promise in identifying favorable alleles in germplasm collections (Yu et al., 2016; Thorwarth et al., 2018).

Upland cotton has a narrow germplasm base, which could allow for elite breeding programs to tap into unexploited standing variation within obsolete cultivars without taking as much of a yield drag associated with crosses to more unadapted germplasm. Developing a training population that represents the U.S cotton obsolete variety collection and phenotyping it using the latest high-quality fiber analysis techniques - HVI, AFIS, and mini-spin will allow breeders to use phenotyping techniques that are much too costly and time consuming to evaluate this collection in its entirety. Genomic prediction will allow for the calculation of predicted phenotypic values within the collection, and the selection of fiber quality alleles which may have been overlooked in the past. The narrowness of the germplasm in this collection is ideal for use in genomic prediction, as relatedness is important in prediction reliability. This would allow breeders to rapidly develop higher fiber quality cultivars that are an important priority for producers.

The following are evaluated in this dissertation: the gains made in cotton fiber quality, the stability of fiber quality traits, and the feasibility and application of genomic prediction for fiber quality. The genetic gains study reports the gains made in HVI and AFIS traits in the last 100 years of US cotton breeding efforts. This will allow for the evaluation of potential unexploited variation for fiber quality traits in the obsolete US cultivar germplasm collection. The stability study of HVI and AFIS traits will ascertain the feasibility of testing a population in limited environments while maintaining the ability to select accurately. The genomic prediction study will provide insights into the use of a modern molecular marker platform in genomic prediction for cotton fiber quality.

1.1. References

Meuwissen, T., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829.

Thorwarth, P., E.A.A. Yousef, and K.J. Schmid. 2018. Genomic prediction and association mapping of curd-related traits in gene bank accessions of cauliflower. G3-Genes Genomes Genetics 8:707-718.

Yu, X., X. Li, T. Guo, C. Zhu, Y. Wu, S.E. Mitchell, K.L. Roozeboom, D. Wang, M.L. Wang, G.A. Pederson, T.T. Tesso, P.S. Schnable, R. Bernardo, and J. Yu. 2016. Genomic prediction contributing to a promising global strategy to turbocharge gene banks. Nature Plants 2:16150.

2. REVIEW OF THE LITERATURE

By the beginning of the 20th century the cotton produced in the United States predominantly originated from accessions brought over in the early 1800s from the Mexican Highlands. Cultivation techniques during the time of introduction were limited in terms of insecticide, and farms were small allowing for fields to be surrounded by abundant natural habits. This resulted in more prolific pollinators, allowing for high rates of cross pollination. Selection of this material was mainly mass selection, as farmers went through their fields and selected bolls from the best plants for the subsequent planting years. Through this form of selection, distinct cultivars began to develop across the United States, and even more so as the art of plant breeding became more defined in the latter part of the nineteenth century (Smith et al., 1999). These early U.S. cultivars were *Gossypium hirsutum*, or upland cotton. Upland cotton, an allotetraploid (2n = 4x = 52), was formed during a polyploidization event 1-2 mya from the joining of two diploid genomes: The A-genome *Gossypium arboreum* with the D-genome *Gossypium raimondii* (Wendel et al., 2009). Although the D-genomes progenitor is non-fiber producing, it is attributed with the majority of fiber quality alleles (Rong et al., 2007). The Mexican Highland introduction showed lots of phenotypic variation; however, U.S. Upland cotton has a very narrow genetic diversity according to molecular studies (Iqbal et al., 2001; Lubbers and Chee, 2009 and references therein).

2.1. Population Structure of U.S. Improved Cultivars

Population structure analyses of U.S. improved cultivars have been conducted using both SSRs and SNPs. Tyagi et al. (2014) reported a structure analysis conducted on 378

cultivars representing 14 cotton producing states, with material releases spanning from 1900 to 2005. This study used 135 SSRs that were developed within the population to best identify the genetic variation. Hinze et al. (2016) reported a structure analysis conducted on 372 U.S. improved cultivars using SSRs. This population was selected to represent four historical and geographical growing regions throughout the United States. This study used 105 SSR markers that were developed by Yu et al. (2012) from a reference panel of various *Gossypium* species to characterize genetic diversity. The most recent study to analyze population structure of U.S. improved cultivars was completed by Hinze et al. (2017), and used SNPs as opposed to the SSRs used in the previous studies. Utilizing the CottonSNP63K array (Hulse-Kemp et al., 2015), this study genotyped 185 U.S. improved cultivars and 26,324 SNPs. In the SSR studies, STRUCTURE (Prichard et al., 2000) was used to evaluate population structure, and in the SNP study, fastSTRUCTURE (Raj et al., 2014) was used. The software fastStructure uses efficient algorithms for approximate inference of the model underlying the STRUCTURE program using a variational Bayesian framework, and was developed to handle large genetic datasets (Raj et al., 2014).

In all studies, genetic variation was limited, but the SSR studies detected significant population structure whereas the SNP study did not. In both SSR studies, five subpopulations were identified that best described the genetic population structure, and were reported to roughly correspond to four geographical U.S. breeding regions: eastern, midsouth, western, and plains. It is interesting that SSRs were able to detect significant population structure and the SNPs were not, considering there are many folds more SNPs used in the analysis than in the SSR analyses. In Maize, SSRs are better at assessing

relatedness than SNPs at similar numbers of markers, but SNPs can be just as informative or more informative when their number exceeds the SSRs by some magnitude (Hamblin et al. 2007, Inghelandt et al. 2010). This is attributed to more polymorphism in an SSR due to more alleles per locus. The question from the cotton population structure studies is: Are the SSRs better at detecting population structure or is the detection of significant structure in the SSR studies the result of a sampling bias due to the small number of SSRs used? In the Tyagi et al. (2014) and the Hinze et al. (2016) study, 135 and 105 SSRs were used respectively, however; the method of development of SSRs was different for each study. As mentioned earlier, the set of SSRs from Tyagi et al. (2014) were developed to characterize genetic variation within the population. This could be seen as bias in the selection of the markers. However, the Hinze et al. (2016) study used SSRs developed from a separate reference population, and both studies showed the same results with the detection of 5 subpopulations. This lends credibility to the argument that the SSRs are better at detecting population structure than SNPs, because even though there over 26,000 more SNPs, the SSRs were able to detect populations structure. However, it is important to note that the SSR studies had twice as many individuals in the analyses as the SNP study. With more individuals in the population, this could lead to better representation and therefore separation of the subpopulations in the analyses. It is reported in the Tyagi et al. (2014) study that only half of the 378 cultivars used could be assigned to the five subgroups at a 70% membership threshold. Hinze et al. (2017) directly compared SSRs to SNPs for a subset of the material. There were 123 cultivars that were genotyped using the 105 SSRs and the 26,324 SNPs. Jaccard similarity matrices were calculated using both marker platforms and compared using

Mantel's correlation r statistic. Mantel's r was determined to by 0.509 indicating a positive but weak relation between the two matrices. The 123 cultivars used in this comparison were made up of 80 U.S. improved cultivars, and the rest were improved *Gossypium hirsutum* cultivars developed in other countries. In order to determine which platform is better at detecting a structure, a more direct analysis will need to be conducted in the future.

2.2. Selection for Fiber Quality

Selection for fiber quality has varied through the years based on resources, and standardization protocols available to the breeder. Prior to standardization, fibers were likely still selected for, but the protocols used were respective to the breeder. It is likely breeders phenotyped fiber quality subjectively in the field, and by physically pulling on fibers to test for strength and judge length; however, these methods were highly subjective, and can be affected by the environment. A history of the standardization of fiber phenotyping methods is described in Ramey (1999). In summary, cotton staple length wasn't standardized until 1918 by the Cotton Futures Act. This was performed by a "pull" test, in which the fiber went through a process of pulling, lapping and discarding to develop a sample that was then measured. Although standardized, this method was still subjective. A mechanical test for fiber strength was implemented by Chandler in 1926. In this method, a combed bundle of fiber was wrapped and broken in a machine that could measure breaking strength. Using a cross sectional measurement of the wrapped fiber bundle and the breaking strength, tensile strength per square inch was determined (Richardson et al., 1937). The "gold standard" of fiber length and length distribution measurements was developed in 1932 called the Suter-Webb sorter. This was a series of combs that allowed a highly skilled technician to slowly

comb out individual fibers and place in a length distribution (Webb, 1932). This was a tedious, slow task that was not efficient for use of screening large amounts of material in breeding programs. A more rapid length measurement technique was developed in 1940 called the fibrogram, which takes a fiber beard, brushes it and passes it through a beam of light, and a sensor measures the light that passes through the sample. This information is utilized in the calculation of fiber length parameters in raw bundle fiber cotton samples (Hertel, 1940). The Stelometer was developed in 1953. This instrument utilizes a sample from a fibrogram beard and places it into a pendulum-style machine that accurately measures both strength and elongation (Hertel, 1953).

The measurement for micronaire was developed in Lord (1956). The limitation of micronaire is the confounding of the variables fineness and maturity. The measurement is Krozney's application of Darcy's Law, with air flow through a bundle of fibers being inversely proportional to the specific surface of individual fibers in the sample, which in turn is directly proportional to maturity and fineness of cotton fibers. Micronaire is used in determining value of cotton, with measurements outside the range of 3.5 – 4.9 units discounted (Ramey, 1999). A low micronaire can be due to immature fibers, which are bad, or fine fibers, which are good, and a high micronaire can be due to mature fibers, which are good, or coarse fibers, which are bad (Hequet et al., 2006).

High Volume Instrumentation (HVI) combined many measurements on a single, prepared bundle of fibers and was developed in 1968, but wasn't widely implemented by either private or public breeding programs until the 1980s. HVI analyzes a bundle of fibers that is taken from each sample of a given weight. The list of traits characterizable by HVI

has evolved over the years, but now often includes length, length uniformity, strength, elongation, color, and micronaire. Advanced Fiber Information System (AFIS) was developed in the 1990s, and measured fiber properties on an individual fiber basis. In this system, individual fibers are blown across a beam of light, and sensors determine the light blocked and the time it take for the fiber to pass the beam. From this, fiber length distributions, seed coat neps, maturity, and fineness can be determined. AFIS doesn't measure strength and elongation as with HVI, but AFIS does give more insight into fiber length properties and separates the confounding variables of micronaire, by measuring both maturity and fineness (Hequet et al., 2006). An advantage to length measurements in AFIS as opposed to HVI is in measuring short fibers. HVI uses a clamp that grabs the fiber sample, and many of the short fibers do not protrude far enough to be detected (Kelly et al., 2015). AFIS sends the fiber through a beam using airflow, therefore not impeding short fiber measurements. AFIS is still considered high-throughput, but does require more time for sample preparation and is more expensive than HVI.

2.3. Genetic Gains in Fiber Quality

The U.S. improved cultivar collection represents over 100 years of breeding efforts. The methods of evaluating and selecting these cultivars based on fiber quality has evolved over time to the current methods of HVI and AFIS. As mentioned, HVI is used widely across every step in the breeding pipelines; however, AFIS is not. It is difficult to look through historical records to look at the gains made in fiber quality before the wide-spread use of HVI to do a comparative analysis, as phenotyping methods have varied and contained a degree of subjectivity. In 1936, Brown reported recognized standard commercial cotton

cultivars at the time. In this report, he identified information on where the cultivar originated, year it was selected, and ranges for both length and lint percent. There is no description of the methods used to identify how the values were determined, but it is likely reported from the breeders. Length was likely determined according to "pull" test as this was common for this time period. The paper reports on 28 upland cultivars that were selected from 1890 to 1931. The lengths range from 19.05 mm to 34.93 mm, and lint percent ranges from 31 to 45. When taking the data from this report and imputing into a linear regression model with year, there is no significant gains in either lint percent or length; however, there is a strong negative correlation between the two traits at -0.56, which is corroborated in a more direct analysis of this time period at -0.45 (Dunlavy, 1923). In general, selection for yield has negative effect on fiber quality (Miller and Rawlings, 1967). In looking at gains over time, Bridge et al. (1971) describes a comparative analysis of cultivars released from 1922 to 1944 with three more recent cultivars released from 1959 to 1966. Cultivars were chosen based on their relative commercial importance at the time they were grown. This study concluded that emphasis for selection was placed on yield during this time period, as the cultivars from 1959 to 1966 performed much better than the older cultivars. Fiber properties such as length and strength were lagging in gains as many of the older cultivars were equivalent or possessed higher fiber quality. Bridge and Meredith (1983) reported a similar study again 12 years later using eleven of the same cultivars from the 1971 study, but adding more recent lines respective to publication, and one older line from 1910 for a total of 17 cultivars. The results were similar to the previous study, but they did report that the 1910 variety had far inferior fiber length and strength. This was attributed to selection for

early maturing cultivars in that time period. Culp and Green (1992) conducted a study comparing both commercial and Pee Dee germplasm spanning from 1945 to 1978 to test for yield and fiber properties gains. Twenty-nine cultivars were chosen based on their performance in Southeastern yield trials and commercial production in South Carolina. They reported similar gains in yield to the Bridge and Meredith (1983) and Bridge et al. (1971) studies. In terms of fiber quality, they did not report overall gains, but did mention they were able to move fiber strength alleles forward in the program along with yield alleles. They attributed this to Beasley's Triple Hybrid (Beasley, 1940). Triple Hybrid has been attributed throughout the literature to breaking some of the negative linkage between strength and yield, and was incorporated into many breeding programs (Green and Culp, 1990; Bowman and Gutierrez, 2003; and Bowman et al., 1996). Although many of these studies report the negative correlations between yield and fiber traits, there are reports of simultaneous gains for both yield traits and fiber quality traits. Bayles et al. (2005) reports gains in yield, length, and strength when looking at 12 cultivars released through the Oklahoma Agriculture Experiment Station from 1918 to 1982, and Schwartz and Smith (2008) reported gains in length and strength from 9 cultivars released from 1905 to 2002. Campbell et al. (2011) described gains in the Pee Dee program using 82 released cultivars from the 1980s to 2001. They report an increase in yield gains, but a decrease in fiber length and strength gains over time. This may be due to the fact that early efforts of the program where directed at increasing fiber quality traits but priority shifted to yield as the program evolved. The negative correlations between the traits are attributed to the decreasing in fiber quality over time. In this study, cultivars were separated into groups representing different breeding

cycles over time. Although it was found that fiber properties decreased over time, it was also found that the rate slowed through different breeding cycles. It was concluded that this supports the idea that negative correlation between fiber quality and yield is due to linkage and not pleiotropy, which supports earlier breeding strategies of intermating and backcrossing to breakup negative linkage to develop favorable genotypes for both traits (Miller and Rawlings, 1967; Meredith, 1977).

As discussed, there are many conflicting reports of gains in fiber properties within the literature. It seems that these studies are subject to the populations used and the time periods evaluated. Many of the studies used a small number of diverse cultivars over varying time periods, and other studies evaluated gains from cultivar releases within specific programs. This makes it difficult to draw comparisons between the studies, and may lead to the differing conclusions, as populations were selected using different phenotyping platforms throughout time. Schwartz and Smith (2008) suggested inconsistencies in gains may be due to the 'unavailability of objective measurement technology combined with the difficulty of integrating these genes into genotypes with other, more valuable traits, and the lack of economic incentive to do so.' Whatever the causes for the inconsistency in gains, the variation and availability of high quality fiber phenotyping platforms leads to the hypothesis that there are potentially unexploited fiber quality alleles in historical U.S. obsolete cultivars.

2.4. Stability of Fiber Quality Traits

There is a large genetic component for fiber quality traits from both HVI and AFIS and indicated by high heritabilities and is well documented (Braden and Smith, 2004; Dabbert et al., 2017; Hugie et al., 2017; Zeng and Bechere, 2017). Many studies have been

conducted to look at the interactions of genetic variation with environmental variation (GxE). This is important, as it determines the strategies a breeder uses for test site evaluation and selection. Geng et al. (1987) conducted a study evaluating 43 cotton cultivars from over 18 years of breeding trails. Fiber phenotype data was taken for different length parameters, uniformity, and strength and was combined in an index to determine a quality score. This study reported that the trait quality score tended to be more stable with newer cultivars, and that as varieties with a higher quality score tend to be more stable. This leads to the idea that breeders can simultaneously improve fiber quality and stability.

Campbell and Jones (2005) described 8 commercial cultivars grown over 4 years at 5 different sites throughout South Carolina for stability. Fiber data was analyzed using HVI. A large GxE for yield and strength, but not for other HVI traits. This study found a similar trend to the Geng et al. (1987) study, and reported that lower performing cultivars regarding strength were more variable across the environmental index. Geng et al. (1987) did not look at strength specifically, but did include strength in the quality score. Campbell and Jones (2005) found that the other HVI traits did not have a large GxE component. This indicates that Geng et al. (1987) result may be due strength.

Campbell et al. 2012 conducted a study looking at 82 released cultivars from the Pee Dee breeding program from 8 different breeding cycles, and analyzed fiber quality using both HVI and AFIS. They report GxE that the proportion of the sum of squares for GxE to the total sum of squares is significant for uniformity, micronaire and fineness. However; when dissecting GxE for the different breeding cycles they report that GxE is more significant for magnitude rather than rank changes, and that quality performance is generally

a stable trait. This is repeated by Ng et al. (2013), who conducted a study looking at the stability of cultivars selected for upper half mean length and strength. They found that these traits were highly stable, and had a high repeatability.

In evaluating these studies, there appears to be potential problem in selecting on obsolete cultivars for fiber strength. According to Geng et al. 1987 and Campbell and Jones (2005), varieties with lower strength have more variation across environments, which could affect values calculated from a combined analysis. This would cause a problem if the values were the desired outcome of the study; however, if selection of top individuals were the desired outcome of a study, then this may be fine as long as the top selected cultivars are of high enough value to be stable.

2.5. Genomic Prediction

An important role of a successful breeder is to identify, create, maintain, and exploit genetic variation through efficient selection platforms to develop improved cultivars. As time has gone by, the tools available to a breeder to accomplish this task have evolved. Traditionally, phenotypic variation and pedigree information are used to infer genetic variation. Henderson (1984) developed mixed model equations for estimation of Best Linear Unbiased Predictors (BLUP) for offspring performance in animal breeding. These models utilize phenotypic data and pedigree information for calculations of genetic variance components, and estimation of breeding values. The base mixed model equation is $\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Zu} + \boldsymbol{e}$, where **y** is a vector of phenotypic values, **X** is a design matrix for fixed effects, **b** is a vector of fixed effects, **Z** is a design matrix for random effects, **u** is a vector of random effects, and **e** is a vector of residuals. Henderson (1984) solution for $\widehat{\boldsymbol{b}}$ and $\widehat{\boldsymbol{u}}$ is the

following: $\begin{bmatrix}\widehat{\boldsymbol{b}}\\ \widehat{\boldsymbol{u}}\end{bmatrix} = \begin{bmatrix}\boldsymbol{X'X} & \boldsymbol{X'Z}\\ \boldsymbol{Z'X} & \boldsymbol{Z'Z}+\boldsymbol{\alpha I}\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{X'y}\\ \boldsymbol{y'Z}\end{bmatrix}$, where **y, X, b, Z,** and **u** are the same as described above with the addition of α, which is the ratio of residual variance and additive genetic variance ($\frac{\sigma_e^2}{\sigma_a^2}$), and **I,** which is an identity matrix with columns and rows equal to the number of random effects. In this formula, the top portion of the equation which solves for the fixed effects is known as the BLUE, which is the best linear unbiased estimate, and the bottom portion of the solution for the random effect is the BLUP. In practice the additive genetic variance and the residual variance are not known, and are estimated using restricted maximum likelihood (REML) (Henderson, 1984). **I** is used in the assumption that all random effects are independent, but in genetics studies, this is not always true as family structure causes correlation. If pedigree information is known, then $\mathbf{A^{-1}}$ is substituted for **I**, where **A** is the pedigree relationship matrix. The variance of genetic effects is var(g) = $\mathbf{A}\sigma_a^2$. BLUPs are used widely in animal breeding as resources are scarcer in terms of number of offspring and time for testing of progeny as opposed to plant breeding. Application of this concept was first used with molecular data in a simulation by Meuwissen et al. (2001), and describes a simulation study used as a proof of concept. The model used for the BLUP was $y = \mu 1_n + \sum_i X_i g_i + e$, where y is a vector of phenotypic values for each individual, μ is the overall mean, 1_n is a vector of ones n individuals in length, X_i is the marker design matrix, g_i is the genetic effects of the markers, and e is the error. In this study σ_e^2 and σ_a^2 are known, and g_i was estimated using mixed models as described by Henderson (1984). This was the beginning of what is now known as genomic prediction.

Early in genomic prediction, pedigree relationship matrices were calculated for use in the mixed model solutions for genetic effects. This was because identity-by-descent (IBD) is necessary to determine the pedigree relationship matrices used in BLUP calculations, and molecular marker data only provides identity-by-state (IBS) information (Isik et al., 2017). Eventually, studies emerged indicating that molecular marker data were effective in approximating LD and pedigree relationships (Habier et al., 2007; Hardy, 2003). VanRaden (2008), developed an efficient algorithm to calculate a genomic relationship matrix (**G**) that is equivalent to **A** from pedigree data using marker data, and is related to the inbreeding coefficient between individual i and j (f_{ij}), where $1\text{-}g_{ij} = f_{ij}$. In this case var(g) is equal to $\mathbf{G}\sigma_a^2$. The use of marker-imputed genomic relationship matrices, has been shown to be more informative at sufficient marker densities than pedigree-based relationship matrices (VanRaden, 2008; Hayes and Goddard, 2008; Engelsma et al., 2012; Albrecht et al., 2014). When the genomic relationship matrix is incorporated into the BLUP it is called the GBLUP.

Another BLUP derivation is the ridge regression BLUP (RR-BLUP), with a base formula of $y = \mathbf{1}_n\mu + \boldsymbol{Wq} + \boldsymbol{e}$, where $\mathbf{1_n}$ is an identity matrix with n (number of observations) rows and columns, μ is the mean, **W** is the genotype matrix, **q** is a vector of the random marker effects, and **e** is a vector of the residual errors. Each column of **W** is coded as 0,1, or 2, then center and standardized by subtracting each element by the $2\ p_j$ which is the minor allele frequency, therefore causing the sum of the column to equal 0 (centered). The genetic effect, and $\hat{\boldsymbol{g}} = [\boldsymbol{W'W} + \lambda\boldsymbol{I}]^{-1}\boldsymbol{W'y}$, where λ is a shrinkage parameter that balances model complexity by goodness of fit. The genetic variance/

covariance matrix is var(g)= **WW'**σ_q^2. Through standardizing and centering, this has been shown to be equivalent to the GBLUP (VanRaden, 2008; Goodard, 2009; Piepho, 2009; Habier et al., 2007; Hayes et al., 2009). The main difference is **G** matrix is calculated to become an n x n matrix, with n being the number of individuals, and **WW'** is a p x p matrix, with p being the number of markers. In genomic prediction, often p is much greater than n; therefore, the GBLUP can be much faster computationally.

A limitation when using BLUPs is the assumption that all markers have equal variance, and marker effects are spread evenly throughout the genome. In order to alleviate this assumption, Meuwissen et al. (2001), developed Bayesian methods (Bayes A and Bayes B) to determine individual marker variances using prior distribution data of the markers themselves. The Bayesian model allows for departure from the infinitesimal model, as loci of large effects can be distinguished from loci with small effects or zero effect (de los Campos et al., 2013). Since the Meuwissen et al. (2001) paper, many different Bayesian models have been implemented by changing the categorization of priors, such as; Bayes Cπ, Bayes Dπ (Lorenz et al., 2011), Bayes Ridge Regression (de los Campos et al., 2009), Bayesian Lasso (Yi and Xu, 2008), and Empirical Bayes (Xu, 2007).

Bayesian methods still assume an additive model, however; not all traits may be additive in nature. A non-parametric model was proposed to better account for nonadditive effects called reproducing kernel Hilbert space RKHS (Gianola and van Kaam 2008). This model uses a square matrix of distances between the observations using the marker data, relaxing the assumption of linearity. Machine learning algorithms such as random forest (RF), support vector machines (SVM), and neural networks (NN), are non-parametric

models that can decipher complex interactions between variables, and is thought to be valuable when looking at traits with large epistatic interactions (Ogutu et al., 2011; Gianola et al., 2011; Howard et al., 2014).

2.6. Comparison of Genomic Prediction Models

There have been many publications comparing models in both simulation and empirical studies. Meuwissen (2009) describes a simulation study where a genome of 10 chromosomes with 1,000 SNPs per chromosome and 12 quantitative trait loci (QTL) were simulated and analyzed using Bayes B and GBLUP. The QTLs were simulated in an additive fashion with differing effect sizes. This study reported that the Bayes B had a higher prediction accuracy determined by Pearson's correlation of predicted value with actual value by 2-6% depending on population size and marker density. It was found that the prediction accuracy for GBLUP increased at a faster rate than the Bayes B model as population size and marker density increased. The fact that Bayes B performed better is expected in this simulation, as the simulation better fit the assumptions of the Bayes B model. The GBLUP distributes all QTL effects evenly, holding more to the infinitesimal model. This simulation had QTLs of differing effects, which is better suited for Bayes B, as Bayesian models loosen the model assumptions of the GBLUP and allows for assignment of QTL with different effects by drawing on different prior distributions for each marker. This biased the study in favor of the Bayes B model. Even though this model performed better, the more rigid GBLUP model was still able to remain within 2-6% of the prediction accuracy.

Daetwyler et al. (2010) performed a similar simulation study, with the modification of changing the number of QTLs contributing to the trait. GBLUP and Bayes B were

directly compared in the simulation study, and again Bayes B performed slightly better than GBLUP as the simulation was created to better fit the Bayes B assumptions. What is added in information is the effect of the different number of QTL that contribute the trait. It was found that as QTL were added, the difference between the prediction accuracies between the models went down. The prediction accuracy for the GBLUP remained relatively the same for all numbers of QTL, but the prediction accuracy for the Bayes B model went down as more QTL were added. Again, this is reasonable, as the GBLUP assumes an infinitesimal model, so adding more QTL approaches this assumption.

Zhong et al. (2009) created simulation that began with empirical data from 42 barely lines. These 42 lines were genotyped with 1,605 markers, and additive QTLs of differing effects and numbers were added. Four different populations were created from the 42 lines: 2 F2 populations from a round robin matting scheme, and two populations randomly mated for five generations. These populations were evaluated for different population sizes. Zhong et al. (2009) reported that the prediction accuracies of the BLUP methods and the Bayesian methods were similar in the F2 populations, and the Bayesian methods did slightly better in the randomly mated populations. However, the BLUP did slightly better when more QTL were added, which concurs with Daetwyler et al. (2010) findings. They then took the four different populations, and randomly mated them for 4 generations. They found that at lower numbers of QTLs that the random mating had less of a decrease in prediction accuracy for the Bayesian models than the BLUP models. Since this simulation began with 42 empirical founder lines. There was likely relatedness between the lines. Even though the QTLs were modeled closer to the assumptions of the Bayesian models, the LD was likely higher in the

original designs. This would have resulted in more correlation between values, which was likely identified in the BLUPs. As lines were further random mated, LD decreased, lowering the correlations and the genetic distances between the individuals and making the genomic relationship matrix less effective on the model. This seemed to allow the Bayesian models to begin to stand out as the simulation better fit the model assumptions.

Howard et al. (2014) compared parametric models to nonparametric models on data sets simulated with entirely additive QTL and entirely epistati QTL. Epistasis was simulated by creating interacting effects of adjacent markers. A broad array of both types of models were used, with BLUP, LASSO, ridge regression Bayes LASSO, the Bayes alphabet (A, B, C, and Cπ) for the parametric models, and Nadaraya-Watson estimator, RKHS, SVM, and NN for the non-parametric models. Prediction accuracies were determined using the correlation of predicted values with actual values. It was found that the parametric models performed generally better than the nonparametric models for the additive QTLs simulation; however, parametric models were not able to predict at all in the entirely epistatic simulations, but the non-parametric models were able to predict a little with prediction accuracies less than 0.4 at a heritability of 0.7 and less than 0.2 at a heritability of 0.3. By alleviate the linear assumptions, non-parametric models have the flexibility to look at the interaction between markers. Since the simulation only had epistatic interactions between QTL, this better suited the non-parametric models.

Iwata and Jannink (2011) compared the non-parametric models RF and SVM to parametric models RR-BLUP, Bayes A, Bayes B, and partial least squares regression. In this study, 863 lines from 9 U.S. barley programs were genotyped using 1,325 SNPs. Due to the

nature of this population, there was strong population structure. 100 additive QTLs were simulated at random maker positions with different heritabilities. Partial least squares was always the worst followed by RF, and all other models were more comparable throughout all heritabilities. This had complicated population structure, but the QTLs were simulated in a simple additive fashion. These results indicate that the population structure didn't allow for contrast between the different models used in this simulated study.

Lorenzana and Bernardo (2009) performed a genomic prediction study on four maize populations and three barley populations using BLUPs and Empirical Bayes (E-Bayes). The populations were between 140 and 339 entries, and differing development methods of RILs, double haploids, F2s from test crosses with RIL populations, and F2s randomly mated for three generations then backcrossed. The markers used were different types and varied in number, with the most being a combination of 1,339 SSRs and RFLPs, and the least being 107 RFLPs. The populations were phenotyped for many traits. They found that the BLUPs were comparable to the E-Bayes, and for many traits was slightly better. These populations had little marker data and relatively small population sizes compared with populations in simulated studies. In the simulated studies mentioned above, the lack of data points would predict that the Bayesian model would perform better. The populations in this analysis are representative of populations that a breeder would be working with for selection, and such were genetically related. This would have given an advantage to the BLUP as indicated by the Zhong et al. (2009) study. Also, the exact genetic architecture of complex quantitative traits is unknown. The Bayesian model is more flexible in this area; however, this was not enough to allow for it to stand out as superior.

Heslot et al. (2012) conducted a genomic prediction study for barley, wheat, and maize in 13 different types of populations. Population sizes ranged from 332 to 761 entries, and the number of markers ranged from 319 to 2,146. This study included non-parametric models of RKHS, SVM, RF, and NN along with the parametric models of RR-BLUP, Bayesian LASSO, Bayesian Shrinkage Regression, Bayes Cπ, and E-Bayes. Again, the RR-BLUP was comparable between all crops, populations and traits, and was the model recommended by the authors. This trend of comparability of BLUPs with other models is consistent throughout the literature with varying crops, populations, and number of markers (Heffner et al., 2011; Spindel et al., 2015; Crossa et al. 2013; Riedelsheimer et al., 2012; Huang et al., 2016). The rigidity of the BLUP appears to hold well under varying circumstances, and due to the rigidity is easy to make inferences. As models move from rigidity to flexibility, the ability to make inferences becomes more difficult, and run a risk of overfitting the model and therefore increasing the mean square error in application (James et al., 2014).

2.7. Predictive Ability

Attempts to understand the predictive ability in genomic prediction have been made in the literature (Daetwyler et al., 2008; Goodard, 2009; Goodard et al., 2011; Lian et al., 2014; Karaman et al., 2016). Goodard et al., 2011 determined the formula: $r^2 \approx q^2 \left(\frac{n_{tr}q^2h^2}{n_{tr}q^2h^2+M_e}\right)$; where r^2 is the squared correlation between the genetic value (μ) and predicted value ($\hat{\mu}$); q^2 is the proportion of genetic variance explained by markers; h^2 is the heritability of the trait; M_e is the effective number of chromosome segments segregating in a population; and n_{tr} is the training population size. This formula identifies the variables that

increase and decrease predictability, but there are implications within these variables that are connected to population structure (Guo et al., 2014, Wientjes et al., 2013, Lorenz and Smith, 2015, etc). According to the formula, increasing M_e will lower predictability; however, M_e is directly affected by population structure. Increasing linkage disequilibrium (LD) in a population will effectively lower M_e concomitantly, lowering the number of markers necessary to capture the informative regions segregating within a population (Elsen, 2016). LD begins to fix regions of the genome through associative mating and inbreeding. This in turn can limit the genetic variation lowering q^2 as well as h^2.

The formula described by Goodard et al. (2011) was rewritten to $R^2 \approx h_M^2 \left(\frac{n_{tr}h_M^2}{n_{tr}h_M^2+M_e}\right)$ (Karaman et al., 2016); the only difference is the terms R^2 which now represents the squared correlation between phenotype (y) and predicted value ($\hat{\mu}$) as μ is not observed in application, and h_M^2 which is the proportion of variance explained by the markers, or the genomic heritability. The change from q^2h^2 to h_M^2 can be made as $h^2 = \frac{\sigma_{\mu}^2}{\sigma_y^2}$, $q^2 = \frac{\sigma_q^2}{\sigma_{\mu}^2}$, and $h_M^2 = \frac{\sigma_q^2}{\sigma_y^2}$. Since $r^2 = \frac{Cov\,(\hat{\mu},\mu)^2}{Var(\hat{\mu})Var(\mu)h^2}$ and $R^2 = \frac{Cov\,(\hat{\mu},\mu)^2}{Var(\hat{\mu})Var(\mu)h^2}$, $r^2h^2 = R^2$. This formula implies that as the training population size is increased, R^2 asymptotically approaches heritability. This is inherently true in that GEBVs are developed from genetic variation.

These formulas do not account for differences among the training population with the validation population due to population structure. When there is imperfect LD between these populations there lies potential genetic variation that is unaccounted for by the reference population in the validation population. For example, if a set of QTLs lie within a

segment of a chromosome that is fixed within the training population due to LD, these QTLs will be amassed as a single value in regards to how they are fixed. If the validation set is segregating within this same region there is added genetic variation unaccounted for in the model. This will result in lower prediction reliability. Another assessment in the Karaman et al. (2016) study was the examination of an upper bound of reliability (UP) originally proposed by Campos et al. (2013). The UP is a limitation to the reliability of prediction based on an imperfect LD between the markers used to compute genomic relationships and QTL. Calculation of the UP is detailed in Karaman et al. (2016). In summary, let $\mathbf{X}_R$ be the genomic relationship matrix of the reference population and R ($\mathbf{X}_R$) be the row space of $\mathbf{X}_R$. Let x_v denote vector of relationships of individual in the validation population with those in the reference population, and is considered the sum of $x_{v1} + x_{v2}$. The components of this summation are further broken down into $x_{v1} = \boldsymbol{Q}_{X_R'} x_v$ and is in R ($\mathbf{X}_R$), and $x_{v2} = (\boldsymbol{I} - \boldsymbol{Q}_{X_R'}) x_v$ and is orthogonal to R ($\mathbf{X}_R$), where $\boldsymbol{Q}_{X_R'} = X_R'(X_R X_R')^{-1} X_R'$. UP of an individual in the validation population is then equal to the ratio of the inner products $\frac{x_{v1}' x_{v1}}{x_v' x_v}$ multiplied by the heritability.

2.8. Genomic Prediction in Germplasm Collections

Recently, genomic prediction has been used successfully in application to germplasm collections with 33,844 photoperiod-sensitive sorghum accessions from the United States Department of Agriculture National Plant Germplasm System (USDA-NPGS) sorghum germplasm collection, consisting of entries from 33 countries and representing five sorghum races (Yu et al., 2016). In this study, authors reported prediction accuracies of 0.76, 0.84,

0.81, 0.75, and 0.90 for biomass yield, dry biomass yield, plant height, root lodging, and stalk number, respectively. The high levels of the prediction accuracies were attributed to optimizing a training population that best represented the germplasm collection. Therefore, increasing the upper bound for reliability. To achieve this, a reference panel of 962 accessions was categorized and a training population was optimized from the results.

Another study on application of genomic prediction on a germplasm collection was conducted in cauliflower (Thorwarth et al., 2018). This study was much smaller in scale than the Yu et al. (2016) study, with only evaluating 174 individuals randomly selected from the collection, but they still showed promising results with prediction accuracies up to .66 depending on the trait. The Yu et al. (2016) study used a GBLUP only, but the Thorwarth et al. (2018) study compared two models; GBLUP and Bayes B. They found that both models had similar prediction accuracies, and could not determine which model was superior in their applications.

The objectives of the study reported herein were to look at the feasibility of using modern genomic tools to select for fiber quality alleles in the obsolete U.S. improved cultivar collection. First, genetic gains of fiber quality traits using modern fiber phenotyping platforms HVI and AFIS was performed to identify if there is potentially untapped genetic variation in this material. Second, stability of fiber quality traits was evaluated to determine if training populations used in genomic prediction needed to be location specific. Third, evaluate the feasibility of using genomic prediction to identify cultivars with potentially beneficial fiber quality

2.9. References

Albrecht, T., H. Auinger, V. Wimmer, J.O. Ogutu, C. Knaak, M. Ouzunova, H. Piepho, and C. Schoen. 2014. Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. Theor. Appl. Genet. 127:1375-1386.

Bayles, M.B., L.M. Verhalen, W.M. Johnson, and B.R. Barnes. 2005. Trends over time among cotton cultivars released by the Oklahoma Agricultural Experiment Station. Crop Sci. 45:966-980.

Beasley, J.O. 1940. The origin of american tetraploid *Gossypium* species. Am. Nat. 74:285-286.

Bowman, D.T., O.L. May, and D.S. Calhoun. 1996. Genetic base of upland cotton cultivars released between 1970 and 1990. Crop Sci. 36:577-581.

Bowman, D.T., and O.A. Gutiérrez. 2003. Sources of fiber strength in the US upland cotton crop from 1980 to 2000. J.Cotton Sci 7:164-169.

Braden, C.A., and C.W. Smith. 2004. Fiber length development in near-long staple upland cotton. Crop Sci. 44:1553-1559.

Bridge, R.R., and W.R. Meredith. 1983. Comparative performance of obsolete and current cotton cultivars. Crop Sci. 23:949-952.

Bridge, R.R., W.R. Meredith, and J.F. Chism. 1971. Comparative performance of obsolete varieties and current varieties of upland cotton. Crop Sci. 11:29-32.

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

Campbell, B.T., P.W. Chee, E. Lubbers, D.T. Bowman, W.R. Meredith Jr., J. Johnson, and D.E. Fraser. 2011. Genetic improvement of the Pee Dee cotton germplasm collection following seventy years of plant breeding. Crop Sci. 51:955-968.

Campbell, B.T., P.W. Chee, E. Lubbers, D.T. Bowman, W.R. Meredith Jr., J. Johnson, D. Fraser, W. Bridges, and D.C. Jones. 2012. Dissecting genotype x environment interactions and trait correlations present in the Pee Dee cotton germplasm collection following seventy years of plant breeding. Crop Sci. 52:690-699.

Crossa, J., Y. Beyene, S. Kassa, P. Perez, J.M. Hickey, C. Chen, G. de los Campos, J. Burgueno, V.S. Windhausen, E. Buckler, J. Jannink, M.A. Lopez Cruz, and R. Babu. 2013. Genomic prediction in maize breeding populations with genotyping-by-sequencing. G3-Genes Genomes Genetics 3:1903-1926.

Culp, T.W., and C.C. Green. 1992. Performance of obsolete and current cultivars and Pee Dee germplasm lines of cotton. Crop Sci. 32:35-41.

Dabbert, T.A., D. Pauli, R. Sheetz, and M.A. Gore. 2017. Influences of the combination of high temperature and water deficit on the heritabilities and correlations of agronomic and fiber quality traits in upland cotton. Euphytica 213:6.

Daetwyler, H.D., B. Villanueva, and J.A. Woolliams. 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. Plos One 3:e3395.

Daetwyler, H.D., R. Pong-Wong, B. Villanueva, and J.A. Woolliams. 2010. The impact of genetic architecture on genome-wide evaluation methods. Genetics 185:1021-1031.

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, and J.M. Cotes. 2012. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 192:1149.

de los Campos, G., A.I. Vazquez, R. Fernando, Y.C. Klimentidis, and D. Sorensen. 2013. Prediction of complex human traits using the genomic best linear unbiased predictor. Plos Genetics 9:e1003608.

Dunlavy, H., T.L. Lyon, J.A. Bizzell, and B.D. Wilson. 1923. Correlation of characters in Texas cotton. Agron. J. 15:444-448.

Elsen, J. 2016. Approximated prediction of genomic selection accuracy when reference and candidate populations are related. Genetics Selection Evolution 48:18.

Engelsma, K.A., R.F. Veerkamp, M.P.L. Calus, P. Bijma, and J.J. Windig. 2012. Pedigree- and marker-based methods in the estimation of genetic diversity in small groups of Holstein cattle. J. Anim. Breed. Genet. 129:195-205.

Geng, S., Q.F. Zhang, and D.M. Bassett. 1987. Stability in yield and fiber quality of California cotton. Crop Sci. 27:1004-1010.

Gianola, D., and van Kaam, Johannes B C H M. 2008. Reproducing Kernel Hilbert Spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178:2289-2303.

Goddard, M.E., B.J. Hayes, and T.H.E. Meuwissen. 2011. Using the genomic relationship matrix to predict the accuracy of genomic selection. J. Anim. Breed. Genet. 128:409-421.

Goddard, M. 2009. Genomic selection: Prediction of accuracy and maximization of long term response. Genetica 136:245-257.

Green, C.C., and T.W. Culp. 1990. Simultaneous improvement of yield, fiber quality, and yarn strength in upland cotton. Crop Sci. 30:66-69.

Guo, Z., D.M. Tucker, C.J. Basten, H. Gandhi, E. Ersoz, B. Guo, Z. Xu, D. Wang, and G. Gay. 2014. The impact of population structure on genomic prediction in stratified populations. Theor. Appl. Genet. 127:749-762.

Habier, D., R.L. Fernando, and J.C.M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389-2397.

Hamblin, M.T., M.L. Warburton, and E.S. Buckler. 2007. Empirical comparison of simple sequence repeats and single nucleotide polymorphisms in assessment of maize diversity and relatedness. Plos One 2:e1367.

Hardy, O.J. 2003. Estimation of pairwise relatedness between individuals and characterization of isolation-by-distance processes using dominant genetic markers. Mol. Ecol. 12:1577-1588.

Hayes, B.J., and M.E. Goddard. 2008. Technical note: Prediction of breeding values using marker-derived relationship matrices. J. Anim. Sci. 86:2089-2092.

Hayes, B.J., P.M. Visscher, and M.E. Goddard. 2009. Increased accuracy of artificial selection by using the realized relationship matrix. Genetics Research 91:143.

Heffner, E.L., J. Jannink, H. Iwata, E. Souza, and M.E. Sorrells. 2011. Genomic selection accuracy for grain quality traits in biparental wheat populations. Crop Sci. 51:2597-2606.

Henderson, C.R. 1984. Applications of linear models in animal breeding. Guelph, Ont. : University of Guelph, 1984.

Hequet, E.F., B. Wyatt, N. Abidi, and D.P. Thibodeaux. 2006. Creation of a set of reference material for cotton fiber maturity measurements. Text. Res. J. 76:576-586.

Hertel, K.L. 1953. The stelometer, it measures fiber strength and elongation. Textile World 103:97-260.

Hertel, K.L. 1940. A method of fibre-length analysis using the fibrograph. Textile Research 10:510-520.

Heslot, N., H. Yang, M.E. Sorrells, and J. Jannink. 2012. Genomic selection in plant breeding: A comparison of models. Crop Sci. 52:146-160.

Hinze, L.L., E. Gazave, M.A. Gore, D.D. Fang, B.E. Scheffler, J.Z. Yu, D.C. Jones, J. Frelichowski, and R.G. Percy. 2016. Genetic diversity of the two commercial tetraploid cotton species in the *Gossypium* diversity reference set. J. Hered. 107:274-286.

Hinze, L.L., A.M. Hulse-Kemp, I.W. Wilson, Q. Zhu, D.J. Llewellyn, J.M. Taylor, A. Spriggs, D.D. Fang, M. Ulloa, J.J. Burke, M. Giband, J. Lacape, A. Van Deynze, J.A. Udall, J.A. Scheffler, S. Hague, J.F. Wendel, A.E. Pepper, J. Frelichowski, C.T. Lawley, D.C. Jones, R.G. Percy, and D.M. Stelly. 2017. Diversity analysis of cotton (*Gossypium hirsutum L.*) germplasm using the CottonSNP63K array. BMC Plant Biology 17:37.

Howard, R., A.L. Carriquiry, and W.D. Beavis. 2014. Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. G3-Genes Genomes Genetics 4:1027-1046.

Huang, M., A. Cabrera, A. Hoffstetter, C. Griffey, D. Van Sanford, J. Costa, A. McKendry, S. Chao, and C. Sneller. 2016. Genomic selection for wheat traits and trait stability. Theor. Appl. Genet. 129:1697-1710.

Hugie, K.L., C.W. Smith, K.S. Joy, and D.C. Jones. 2017. Divergent selection for fiber length and bundle strength and correlated responses in cotton. Crop Sci. 57:99-107.

Hulse-Kemp, A.M., J. Lemm, J. Plieske, H. Ashrafi, R. Buyyarapu, D.D. Fang, J. Frelichowski, M. Giband, S. Hague, L.L. Hinze, K.J. Kochan, P.K. Riggs, J.A. Scheffler, J.A. Udall, M. Ulloa, S.S. Wang, Q. Zhu, S.K. Bag, A. Bhardwaj, J.J. Burke, R.L. Byers, M. Claverie, M.A. Gore, D.B. Harker, M.S. Islam, J.N. Jenkins, D.C. Jones, J. Lacape, D.J. Llewellyn, R.G. Percy, A.E. Pepper, J.A. Poland, K.M. Rai, S.V. Sawant, S.K. Singh, A. Spriggs, J.M. Taylor, F. Wang, S.M. Yourstone, X. Zheng, C.T. Lawley, M.W. Ganal, A. Van Deynze, I.W. Wilson, and D.M. Stelly. 2015. Development of a 63K SNP array for cotton and high-density mapping of intraspecific and interspecific populations of *Gossypium spp.* G3-Genes Genomes Genetics 5:1187-1209.

Iqbal, M.J., O. 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. Theor. Appl. Genet. 103:547-554.

Isik, F., J.B. Holland, and C. Maltecca. 2017. Genetic data analysis for plant and animal breeding. Cham, Switzerland : Springer, 2017.

Iwata, H., and J. Jannink. 2011. Accuracy of genomic selection prediction in barley breeding programs: A simulation study based on the real single nucleotide polymorphism data of barley breeding lines. Crop Sci. 51:1915-1927.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2014. An introduction to statistical learning: With applications in R. corrected edition. New York : Springer, 2014; Corrected edition, .

Karaman, E., H. Cheng, M.Z. Firat, D.J. Garrick, and R.L. Fernando. 2016. An upper bound for accuracy of prediction using GBLUP. Plos One 11:e0161054.

Kelly, B., N. Abidi, D. Ethridge and E.F. Hequet. 2015. Fiber to fabric. p. 665-744. *In* D.D. Fang, and R.G. Percy (eds.) Cotton 2nd edition. American Society of Agronomy, Inc., Crop Science Society of America, Inc., and Soil Science Society of America, Inc, Madison, WI.

Lian, L., A. Jacobson, S. Zhong, and R. Bernardo. 2014. Genomewide prediction accuracy within 969 maize biparental populations. Crop Sci. 54:1514-1522.

Lord, E. 1956. 2—Air flow through plugs of textile fibres: Part II. the micronaire test for cotton. Journal of the Textile Institute Transactions 47:T16-T47.

Lorenz, A.J., and K.P. Smith. 2015. Adding genetically distant individuals to training populations reduces genomic prediction accuracy in barley. Crop Sci. 55:2657-2667.

Lorenz, A.J., S. Chao, F.G. Asoro, E.L. Heffner, T. Hayashi, H. Iwata, K.P. Smith, M.E. Sorrells, and J. Jannink. 2011. Genomic selection in plant breeding: Knowledge and prospects. Advances in Agronomy, Vol 110 110:77-123.

Lorenzana, R.E., and R. Bernardo. 2009. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor. Appl. Genet. 120:151-161.

Lubbers, E.L., and P.W. Chee. 2009. The worldwide gene pool of *G. hirsutum* and its improvement. p. 23-52. *In* A. Patterson (ed.) Genetics and genomics of cotton. Springer, New York.

Meredith, W.R. 1977. Backcross breeding to increase fiber strength of cotton. Crop Sci. 17:172-175.

Meuwissen, T., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829.

Meuwissen, T. 2009. Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. Genetics Selection Evolution 41:35.

Miller, P.A., and J.O. Rawlings. 1967. Selection for increased lint yield and correlated responses in upland cotton *Gossypium hirsutum L.* Crop Sci. 7:637-640.

Ng, E., K. Jernigan, W. Smith, E. Hequet, J. Dever, S. Hague, and A.M.H. Ibrahim. 2013. Stability analysis of upland cotton in Texas. Crop Sci. 53:1347-1355.

Ogutu, J.O., H. Piepho and T. Schulz-Streeck. 2011. A comparison of random forests, boosting and support vector machines for genomic selection. p. S11. *In* A comparison of random forests, boosting and support vector machines for genomic selection. BMC proceedings, 2011. BioMed Central.

Piepho, H.P. 2009. Ridge regression and extensions for genomewide selection in maize. Crop Sci. 49:1165-1176.

Pritchard, J.K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945-959.

Raj, A., M. Stephens, and J.K. Pritchard. 2014. Variational inference of population structure in large SNP datasets. Genetics 197:573-589.

Ramey Jr, H.H. 1999. Classing of fiber. p. 709-727. *In* C.W. Smith, and J.T. Cothern (eds.) Cotton: Origin, history, technology, and production. John Wiley & Sons, New York.

Richardson, H.B., T.L. Bailey, and C.M. Conrad. 1937. Methods for the measurement of certain character properties of raw cotton. US Department of Agriculture.

Riedelsheimer, C., F. Technow, and A.E. Melchinger. 2012. Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines. BMC Genomics 13:452.

Rong, J., E.A. Feltus, V.N. Waghmare, G.J. Pierce, P.W. Chee, X. Draye, Y. Saranga, R.J. Wright, T.A. Wilkins, O.L. May, C.W. Smith, J.R. Gannaway, J.R. Wendel, and A.H. Paterson. 2007. Meta-analysis of polyploid cotton QTL shows unequal contributions of subgenomes to a complex network of genes and gene clusters implicated in lint fiber development. Genetics 176:2577-2588.

Schwartz, B.M., and C.W. Smith. 2008. Genetic gain in fiber properties of upland cotton under varying plant densities. Crop Sci. 48:1321-1327.

Smith, C.W., R.G. Cantrell, H.S. Moser and S.R. Oakley. 1999. History of cultivar development in the United States. p. 99-173. *In* C.W. Smith, and J.T. Cothren (eds.) Cotton: Origin, history, technology, and production. John Wiley & Sons, New York.

Spindel, J., H. Begum, D. Akdemir, P. Virk, B. Collard, E. Redona, G. Atlin, J. Jannink, and S.R. McCouch. 2015. Genomic selection and association mapping in rice (*oryza sativa*): Effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. Plos Genetics 11:e1004982.

Thorwarth, P., E.A.A. Yousef, and K.J. Schmid. 2018. Genomic prediction and association mapping of curd-related traits in gene bank accessions of cauliflower. G3-Genes Genomes Genetics 8:707-718.

Tyagi, P., M.A. Gore, D.T. Bowman, B.T. Campbell, J.A. Udall, and V. Kuraparthy. 2014. Genetic diversity and population structure in the US upland cotton (*Gossypium hirsutum L.*). Theor. Appl. Genet. 127:283-295.

Van Inghelandt, D., A.E. Melchinger, C. Lebreton, and B. Stich. 2010. Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. Theor. Appl. Genet. 120:1289-1299.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414-4423.

Webb, R.W. 1932. The suter-webb cotton fiber duplex sorter and the resulting method of length-variability measurements. Am.Soc.Test.Mater.Proc.32 (2)764-771.

Wendel, J.F., C. Brubaker, I. Alvarez, R. Cronn and J.M. Stewart. 2009. Evolution and natural history of the cotton genus. p. 3-22. *In* A. Patterson (ed.) Genetics and genomics of cotton. Springer, New York.

Wientjes, Y.C.J., R.F. Veerkamp, and M.P.L. Calus. 2013. The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. Genetics 193:621-631.

Xu, S. 2007. An empirical Bayes method for estimating epistatic effects of quantitative trait loci. Biometrics 63:513-521.

Yi, N., and S. Xu. 2008. Bayesian LASSO for quantitative trait loci mapping. Genetics 179:1045-1055.

Yu, J.Z., D.D. Fang, R.J. Kohel, M. Ulloa, L.L. Hinze, R.G. Percy, J. Zhang, P. Chee, B.E. Scheffler, and D.C. Jones. 2012. Development of a core set of SSR markers for the characterization of *Gossypium* germplasm. Euphytica 187:203-213.

Yu, X., X. Li, T. Guo, C. Zhu, Y. Wu, S.E. Mitchell, K.L. Roozeboom, D. Wang, M.L. Wang, G.A. Pederson, T.T. Tesso, P.S. Schnable, R. Bernardo, and J. Yu. 2016. Genomic prediction contributing to a promising global strategy to turbocharge gene banks. Nature Plants 2:16150.

Zeng, L., and E. Bechere. 2017. Correlated selection responses of fiber properties measured by high volume instrument and advanced fiber information system in upland cotton. Euphytica 213:278.

Zhong, S., J.C.M. Dekkers, R.L. Fernando, and J. Jannink. 2009. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: A barley case study. Genetics 182:355-364.

3. GENETIC GAINS OF COTTON FIBER QUALITY IN THE PAST 100 YEARS

3.1. Introduction

The predominant genetic base of cotton grown in the United States can be traced from accessions brought over in the early 1800s from the Mexican Highlands (Lubbers and Chee, 2009). The highland introduction created a genetic bottleneck resulting in narrow genetic diversity according to molecular studies (Iqbal et al., 2001; Tyagi et al. 2014, Hinze et al. 2015). Tyagi et al. 2014 concluded that the low genetic diversity could allow for elite breeding programs to tap into unexploited standing variation within obsolete cultivars without suffering a yield drag as associated with crosses to more wild germplasm.

Since the Mexican introduction, farmers and breeders have made selections for fiber quality using the resources and techniques available at that time. These selection methods began with subjective field evaluations, evolving to quantifiable techniques that were slow and tedious and therefore not widely used throughout selections programs (See Chapter 1). In more recent years, high quality cotton fiber phenotyping methods allow for high-throughput evaluation of bundle fiber samples through HVI, and individual fiber samples such as AFIS. Many of the traits measured by HVI and AFIS were not previously measured and therefore not selected on prior to their incorporation into breeding programs.

The US National Cotton Germplasm Collection currently maintains 6,302 upland cotton accessions consisting of 2,522 landraces and 3,780 improved cultivars (Campbell et al., 2010). The 3,780 improved cultivars represent over a hundred years of selection efforts for cotton improvement. As an inbreed crop, this collection also offers a snapshot of the

genetic merits of that time period and growing region. With the low differences in genetic diversity between elite breeding material and the fact that past selection efforts in cotton fiber quality were limited, this study looks to evaluate if genetic variation has been tapped into over the past 100 years of breeding efforts in cotton fiber traits. In this study a population was developed to represent 100 years of breeding efforts from different growing regions in the US, and a genetic gains analysis was conducted to determine gains in HVI and AFIS traits.

3.2. Materials and Methods

Tissue samples were collected from young leaves, and DNA was extracted using a modified CTAB (cetytrimethylammonium bromide) method described by Zhang et al. (2010). A collection of cultivars from the USDA U.S. improved cotton cultivar collection was genotyped using Illumina® 63K SNP array (Hinze et al., 2015; Hulse-Kemp et al., 2015). This population was developed to represent historical and geographically distinct breeding efforts in the United States. SNP markers were removed 1) when markers were non-polymorphic, 2) greater than 10 % of SNP calls were missing in population, 3) minor allele frequency was less than .03, and 4) heterozygosity of marker was greater than 10 %. After marker filtering, 20,491 high quality SNPs remained. Genetic diversity was analyzed by calculation of Identity by state (IBS) matrix, and performing a Principle Coordinate Analysis (PCoA) in R (R core team 2016). Individual cultivars were selected for genetic gains analysis that genetically best represented the U.S. improved cotton cultivar collection and breeding efforts over time from 1900-2015. Sixty-three cultivars were chosen based on these criteria (GGpop).

GGpop was planted in a randomized complete block design (RCBD) in 2016 and 2017 in Weslaco, TX at the Texas A&M AgriLife Research and Extension Center and in Corpus Christi, TX at the Texas A&M AgriLife Research and Extension Center. In 2016, three replications were used at both locations, and two replications in 2017 were used at both locations. Soil type at Weslaco is a Hidalgo sandy clay loam, a fine-loamy, mixed, active, hyperthermic Typic Calciustolls, and at Corpus Christi a Houston black clay, a fine smectitic, thermic Udic Haplustert. Normal cotton production practices were used in all trails, with furrow irrigation used in Weslaco, TX. Boll samples were randomly harvested from plots with 30 bolls taken from the first fruiting limb position in the middle of the fruiting zone. Fiber was ginned using 8-saw laboratory gins, with each rep ginned by a single gin, and fiber phenotyping was performed using High Volume Instrumentation (HVI) and Advanced Fiber Information System (AFIS) at the Fiber and Biopolymer Research Institute at Lubbock, TX. Traits analyzed are listed in Table 3.1. Phenotype data were analyzed with mixed linear models using lme4 package in R (Bates et al., 2015), and Empirical Best Linear Unbiased Predictors (EBLUPs) were calculated from models for each accession for each trait. EBLUPs were then regressed on years that cultivars were released to calculate genetic gains using the lm function in r.

3.3. Results and Discussion

All points in the PCoA depicted in Figure 3.1 are U.S. improved cultivars, and the black points represent the cultivars selected for this study. The distributions in Figure 3.1 indicates that the cultivars selected for this study are well distributed and a good

representation of the genetic space of the U.S. improved cotton cultivar collection. Cultivars chosen for this genetic gain study are listed in Table 3.2.

The gains for all traits are listed in Table 3.3. Lint % is a yield component, and is used as a gauge of yield (Miller and Rawlings 1967). Lint% showed significant gains over the years at 0.174 % increase per year and a $p < 0.001$. The R^2 was 0.42, which was the highest for any trait, showing the strongest relationship with year compared to any other trait. Yield is often the highest selection priority for breeding programs, and the strength of this model compared to others offers evidence of this.

Significant gains have been made in creating longer fibers. This is seen in gains for UHML, UQL(w), L5(n), L(w), and L(n) at 0.094, 0.098, 0.087, .103, .111 percent gain per year, respectively (Table 3.3), and all gains are significant with $p <0.001$. Increases are being made in the mean fiber length, as well as the longest fibers in a sample. These traits are also strongly positively correlated. UHML has a higher correlation with AFIS mean length based on fiber weight, L(w), at .96 than AFIS mean length based on actual length measurements regardless of fiber weight, L(n), at .85. This is likely due to UHML being a measurement taken by weight also. The significant gains show that the various selection methods have been effective for these traits.

Significant gains are also seen in reducing the short fiber content in samples. The short fiber content both by weight and length have shown to decrease per year (Table 3.3). It is interesting to note that percent decrease in SFC(w) is faster than the decrease in SFC(n) with a rate of -0.267 % and -0.146 % reduction in SFC per year, respectively. Both of these gains are significant with $p > 0.001$. It is not expected that selection intensity is greater for

SFC(w), or that direct selection for SFC in either form was performed much at all over the past 100 years of breeding efforts. Sutter-Webb method was developed in 1932 (Webb, 1932), which allows for selection against SFC is too tedious to use as a selection platform. AFIS allows for selection against SFC, but wasn't developed until the 1990s and is not widely utilized by breeding programs even today. Selection against SFC is likely a product of negative correlation with length measurements. The greater negative gains in SFC(w) over SFC(n) is a result in bias created in length measurements. SFC(w) has a stronger negative correlation with length measurements than SFC(n). Length by weight measurements calculate mass from an assumed uniform density, but not all fibers have the same density. Correlations in this study show that Maturity increases as length traits increase, meaning more secondary cell wall development in the fiber, thus denser fibers (Hequet et al. 2006). This causes an over prediction of longer fibers by weight, and an under prediction of SFC(w). Since, fibers have been getting longer over time through selection, the negative correlation of short fiber content is compounded by the under prediction of SFC(w). This gives the appearance that SFC(w) has a greater reduction in gain than SFC(n).

Gains in the uniformity of length data are less clear than gains in length and gains in reducing short fiber content. There is a significant gain in Unif at 0.025 % gain per year which is significant at $p < 0.001$. The term uniformity as used in HVI (Unif) is misleading. The true uniformity of a length distribution should incorporate the short fibers in the measure and HVI doesn't do a good job of calculating short fibers (Krowicki and Ramey 1984). Unif is the measure of mean length by weight divided by the UHML multiplied by 100. An increase in Unif would be the result in a greater rate of gain in mean length by

weight than UHML, or a decrease in UHML which would not be desirable. Since gains in AFIS percent L(w) are greater than UHML and L(w) is a similar measurement to the mean length that is calculated by HVI, then a better measure of uniformity is using the coefficients of variation for the fiber lengths within the sample (Table 3.3). The coefficients of variation for both L(w) and L(n) are determined by dividing the standard deviation by the mean length. The only difference between these traits is how mean length is determined, either by weight or by number. For these traits, breeders have improved L(w)CV by -0.044, $p < 0.01$, but L(n)CV of -0.02 was not significant, $p = 0.345$. The gain in L(w)CV is likely the product of the measurement. The mean length by weight is over-predicted, and as a result increases at a faster rate than L(n) (Figure 3.3). As breeders have selected for longer fibers over the past 100 years, the HVI measurement for mean length by weight, which is the denominator of the formula for L(w)CV, is over-predicted because the internal algorithm assumes uniform weight throughout the length of the fiber. This would result in a sharper decrease in L(w)CV according to the calculation of coefficient of variation. As L(n) is not biased by the assumption of uniform density, it is not affected by the increase in length over time, and therefore doesn't affect the L(n)CV measurement over time. This concept was further explored by regressing L(w) to L(w)CV and L(n) to L(n)CV. Since L(n)CV has more variation than L(w)CV, the values were centered and scaled for direct comparison. L(w)CV did decrease at a higher rate as L(w) increased than L(n)CV as L(n) increased. The regression coefficient for L(w)CV/L(w) was -3.909 and the regression coefficient for L(n)CV/L(n) was -2.675, and the p-values for the coefficients were .089 and .332,

respectively. Thus, L(w), the denominator of L(w)CV, had a larger effect on the measure than L(n), the denominator of L(n)CV.

HVI strength has a strong positive correlation with fiber length parameters, and also shows a positive gain of 0.157 % per year. When plotting the gains for HVI strength it was noted that data didn't appear to have an entirely linear growth trend (Figure 3.4). To explore this, polynomial regression lines were fitted at 2,3, and 4 degrees, and polynomial models and the linear regression model were compared using the Akaike's Information Criterion (AIC) to determine the best model. The AIC scores for the linear, 2^{nd}, 3^{rd}, and 4^{th} degree polynomial equations were 261.17, 260.15, 262.13, and 262.13 respectively. The lowest AIC was for the 2^{nd} degree polynomial equation indicating it was the best model in explaining the data, and is represented by the green line in Figure 3.4. This model best represents the data numerical, but also intuitively. Looking at the polynomial regression line, there appears to be a sharp increase in gains starting around the 1940s. This is the time that Beasley's Triple Hybrid was developed, which was rapidly integrated into breeding programs and is attributed to breaking the negative linkage between strength and yield (Beasley 1940, Green and Culp 1990, Bowman and Gutierrez 2003, and Bowman et al. 2006). A few years later the invention of the Stelometer (Hertel 1953), gave breeders the ability to phenotype relatively efficiently for cotton fiber bundle strength. HVI Strength and Elongation have been reported to have negative correlation (May and Taylor 1998), and this holds true in this study as well, with a correlation of -0.38. In general, selection for elongation among breeders is not practiced (Benzina et al. 2007), which could explain the results of this study

in which elongation was found to have significantly reduced by -0.102 % per year with $p < 0.01$.

MIC is a trait that is confounded by maturity and fineness (Hequet et al. 2006). It is normally selected to remain within a range, as values of 3.7-4.2 will receive a premium price, and under 3.4 and over 5.0 will be penalized with a discount. As expected, there was no significant linear relationship with year, as selection pressure maintains this range. However, the major components of MIC, maturity and fineness, have seen significant gains. Mat. Ratio and IFC had a strong negative correlation of -0.95. Both of these traits showed significant gains ($p < 0.001$) with a positive gain of 0.064 % per year with maturity, and a negative gain of -0.276 % per year for IFC. Fine did not show significant gains with $p = 0.1687$; however, Std. Fine did show a significant negative gain of 0.064 % per year with $p < 0.001$. Since Std. Fine is simply fineness divided by maturity, this gain is attributed mainly to the gain in maturity. Of these traits, the only one likely directly selected upon over the past 100 years is micronaire, as the test for micronaire was developed in 1956 (Lord 1956), but as mentioned, did not show significant gains. The significant gains made in Mat. Ratio and IFC are likely the result of indirect selection pressure. The traits that have the strongest correlations with Mat. Ratio and IFC, have made significant gains, and were directly selected upon, were Strength and Lint%. Strength seems the most reasonable contributor to these gains, as maturity is determined by secondary wall development in the fiber. As more cellulose is created, Strength should increase. Also, Strength is tested as a bundle of fibers, so the less IFC, the stronger the fiber bundle. If gains for these traits were the result of selection to increase Strength, it is expected to see a similar trend as Strength over time as

mentioned earlier. Polynomial regressions were run the same for Mat. Ratio and IFC as Strength. In both cases the lowest AIC was for the 2nd degree polynomial regression model. The same trend can be seen in Figure 3.4 for both Mat. Ratio and IFC as was displayed for Strength. Around the 1940s the rate greatly changes, with Mat. Ratio increasing, and with IFC decreasing. Indicating that improvements in these area are likely due to improvements in strength.

3.4. Conclusion

The population used in this study is a good representation of the material from the US National Cotton Germplasm Collection's obsolete improved cultivar collection according to the marker platform used in this analysis. There have been steady gains in increasing fiber length parameters for both HVI and AFIS, and reducing short fiber content. HVI Unif has shown significant gains, but the coefficient of variation for length by number, which is the preferred measurement of uniformity by the author has not seen significant gains. There were significant gains seen in strength, and there appears to be a sharp increase in gains starting around the 1940s, which is consisted with the hypothesis that the development of Beasley's Trible Hybrid contributed to the breaking of negative linkage with yield. Elongation has seen a significant decrease over the past hundred years, and is attributed to the negative linkage with strength. There have been significant gains in Mat. Ratio, and standard fineness, and reducing IFC; however, selection has not been directly used on these traits. It is likely that gains in these traits are from correlation with traits such as length and strength in which there has been direct selection pressure. Since correlation doesn't explain all of the variation for these traits, it is probable that there is much untapped

genetic variation for these traits in the obsolete US improved material. Breeders may find value in looking to this collection to improve these fiber traits.

Table 3.1. Traits analyzed in genetic gains study. Sixty-three obsolete and near modern cultivars from the U.S. Cotton Germplasm Collection were grown at Corpus Christi and Weslaco, TX in 2016 and 2017.

System	Symbol	Trait	Unit	Description
Scale	Lint %	Lint Percent	%	Percent of seed cotton weight that is fiber weight
HVI	MIC	Micronaire	Unitless	Test for fineness and maturity using relationship between airflow and linear density
	UHML	Upper Half Mean Length	mm	Mean of the longest 50% of fibers
	Unif	Length Uniformity	%	The ratio between the mean length and the upper half mean length.
	Strength	Bundle Strength	Kn x m/kg	Force to break a bundle of fiber
	Elon	Elongation	%	The percentage of change in a bundle length before rupture under a breaking load
AFIS	L(w)	Length by Weight	mm	The mean length of the sample by weight
	L(w)CV	Length variation by weight	%	A measure of the standard deviation of the fiber length within a sample by weight standardized by the average fiber length
	UQL(w)	Upper Quartile Length by Weight	mm	The length that is exceeded by 25% of the fibers by weight
	SFC(w)	Short Fiber Content by Weight	%	The percentage of fibers by weight that are shorter than 12.7mm in length
	L(n)	Length by Number	mm	The mean length of the sample by number
	L(n)CV	Length variation by number	%	A measure of the standard deviation of the fiber length within a sample by weight standardized by the average fiber length
	SFC(n)	Short Fiber Content by number	%	The percentage of fibers by number that are shorter than 12.7mm in length
	L5(n)	Length Exceeded by 5% of all fibers	mm	The Length that is exceeded by the longest 5% of the fibers in the sample based on the length-by number distribution.
	Fine	Fineness	mtex	The linear density of fiber defined as mass per unit length
	IFC	Immature Fiber Content	%	Percentage of fibers with less than 0.25 degree of wall thickening
	Mat. Ratio	Maturity Ratio	%	Percent of fibers greater than 0.5 degree of wall thickening minus IFC divided by 200 and added to 0.7
	Std. Fine	Standard Fineness	No unit	Fineness standardized by the maturity ratio

Table 3.2. Accessions used in this calculating genetic gains.

Name	Year	Source for determining Year
Mebane	1897	Smith et al., 1999
Lone Star	1904	Brown, 1936
Durango	1905	Smith et al., 1999
Hartsville	1905	Bowman et al., 2006
Half and Half	1906	Brown, 1936
Meade Clean Seed	1912	Brown, 1927
Express-432	1914	Brown, 1927
Dixie Triump	1915	Brown, 1936
Deltatype Webber	1915	Smith et al., 1999
Cleveland W.R. Wannamaker's	1916	Brown, 1927
New Boykin	1918	Brown, 1936
Lightning express	1923	Brown, 1936
Coker's Clevewilt 3	1932	Bowman et al., 2006
Coker 100 wilt	1941	Smith et al., 1999
Deltapine 14	1941	Bowman et al., 2006
Bobshaw 1	1941	Bowman et al., 2006
Lankart 57	1950	Okelly, 1950
Lockett 1	1950	Arnold, 1975
Western Stormproof	1950	Bowman et al., 2006
Auburn 56	1953	Smith et al., 1999
Dixie King	1956	Smith et al., 1999
Blight Master	1956	Ramey, 1966
Fox 4	1958	Ewing, 1965
Stoneville 213	1962	Bowman et al., 2006
Del Cerro	1962	Bowman et al., 2006
Pope	1964	Duncan and Pate, 1964
Deltapine 16	1968	Jones, 1998
Westburn	1969	Smith et al., 1999
Delcot 277	1970	Sappenfield et al., 1972
Coker 310	1971	PVP-7100021
Coker 312	1972	PVP-7200100
Stoneville 256	1975	PVP-7500102
Deltapine 55	1975	PVP-7500103
Acala SJ-5	1977	Bowman et al., 2006
Cascot L-7	1977	PVP-7700043

Table 3.2. Continued.

Name	Year	Source for determining Year
DES 56	1978	PVP-7800041
McNair 235	1978	PVP-7800068
Dunn 219	1978	PVP-7900006
PD 2165	1979	Harrell and Culp, 1979
Stoneville 825	1979	PVP-7900024
Paymaster 145	1980	PVP-8000080
Earlistaple 7	1980	Culp and Harrel, 1980
Deltapine 90	1984	PVP-8100143
DP 50	1984	PVP-8400154
PD-2	1985	Culp et al., 1985a.
DES 119	1985	PVP-8500176
PD1	1985	Culp et al., 1985b
HS 26	1986	PVP-8600087
Acala Maxxa	1990	PVP-9000168
Paymaster HS 200	1990	PVP-9000216
DPL 5690	1991	PVP-9100116
Georgia King	1991	PVP-9100257
All-Tex Atlas	1992	PVP-9200188
Ciano Cocorim 92	1992	Jasso and Solis, 1994
MD51ne	1993	Meredith, 1993
LA887	1993	PVP-9100065
Acala 1517-99	2000	Cantrell et al., 2000
PSC 355	2000	McPherson et al., 2000
Arkot A306	2000	Bourland and Smith, 2001
DPL 491	2001	PVP-200100159
Phytogen 72	2001	PVP-200100115
Tamcot Sphinx	2001	PVP 9600134
UA48	2010	PVP-201100041
Tamcot 73	2011	Smith et al., 2011
Commercial	2015	Commercial company

Table 3.3. Gains per year in respective unit and as percent.

Traits	(Unit) Gains/Year	(%) Gains/Year	$R^{2\dagger}$	Correlation‡
Lint%	0.1738***	0.174***	0.4211	0.6489
Elon	-0.1021**	-0.102**	0.1552	-0.3939
Strength	0.1572***	0.157***	0.3419	0.5848
UHML	0.0940***	0.094***	0.2226	0.4718
Unif	0.0251***	0.025***	0.3041	0.5515
UQL(w)	0.0976***	0.098***	0.2514	0.5014
L5(n)	0.0872***	0.087***	0.2347	0.4845
L(w)	0.1028***	0.103***	0.3032	0.5506
L(n)	0.1108***	0.111***	0.3531	0.5943
IFC	-0.2756***	-0.276***	0.3137	-0.5601
SFC(w)	-0.2673***	-0.267***	0.2887	-0.5373
SFC(n)	-0.1459***	-0.146***	0.1808	-0.4252
L(w)CV	-0.0439**	-0.044**	0.1205	-0.3471
L(n)CV	-0.0201	-0.02	0.0146	-0.1209
MIC	0.0500	0.05	0.0422	0.2053
Mat. Ratio	0.0645***	0.064***	0.3848	0.6203
Fine	0.0282	0.028	0.0308	0.1756
Std. Fine	-0.0331*	-0.033*	0.0644	-0.2538

† Coefficient of Determination for the gains model

‡ Pearson Correlation Coefficient for the trait with year.

***, **, * Significance at $p < 0.001$, $p < 0.01$, and $p < 0.05$ respectively

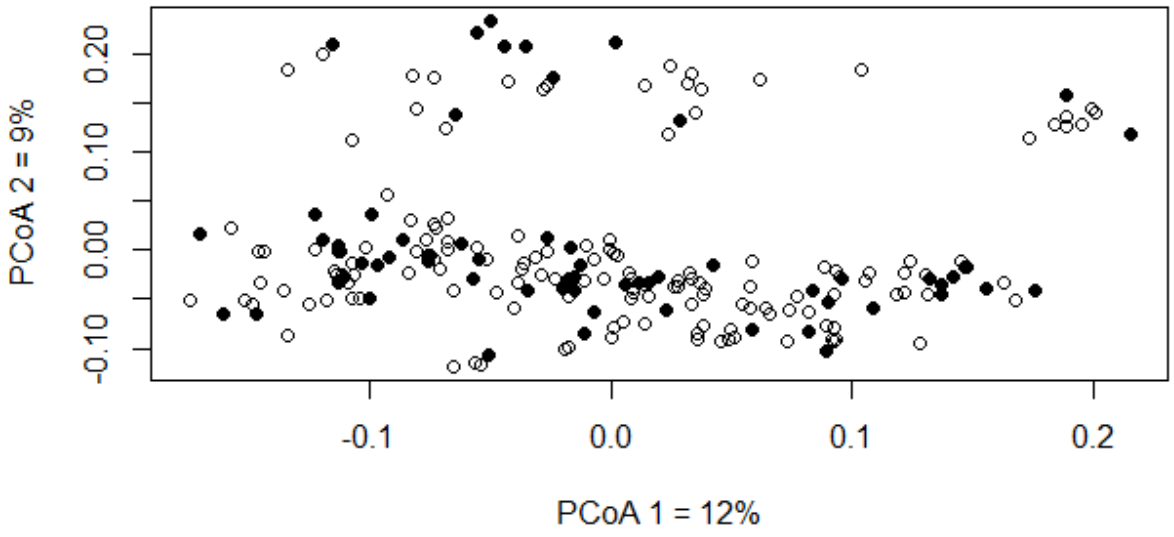


Figure 3.1. Plot of 1st and 2nd Principle Coordinates from Principle Coordinate analysis of the USDA's U.S. improved cotton cultivar collection. The black dots indicate the 63 accessions used in this genetic gains analysis.

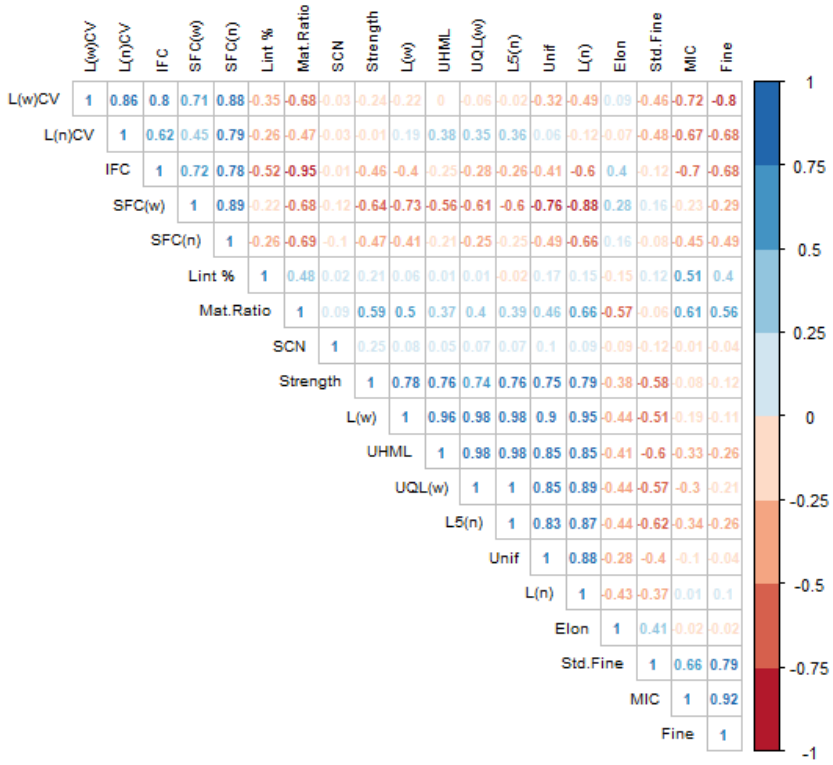


Figure 3.2. Correlations among all traits analyzed for genetic gains. The brighter colors indicate a strong positive correlation, the lighter colors indicate a weaker correlation, the red colors indicate negative correlation, and the blue colors indicate positive correlation.

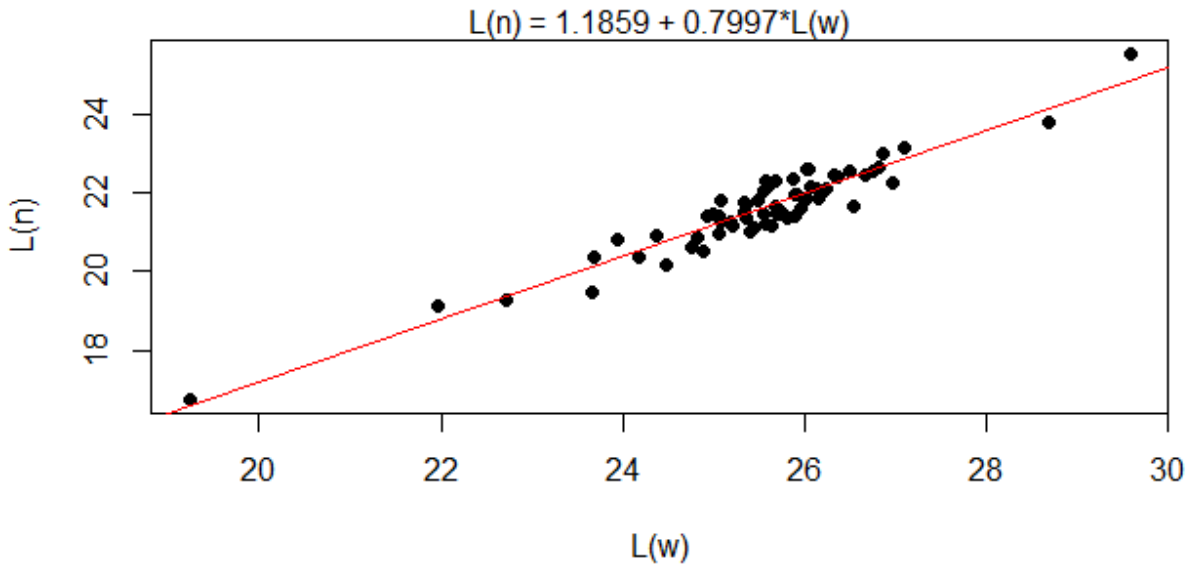


Figure 3.3. Plot of L(w) by L(n). For every 1 mm increase in L(w), L(n) only increases by 0.7997 mm. The red line represents the regression line from the formula on the top of the plot.

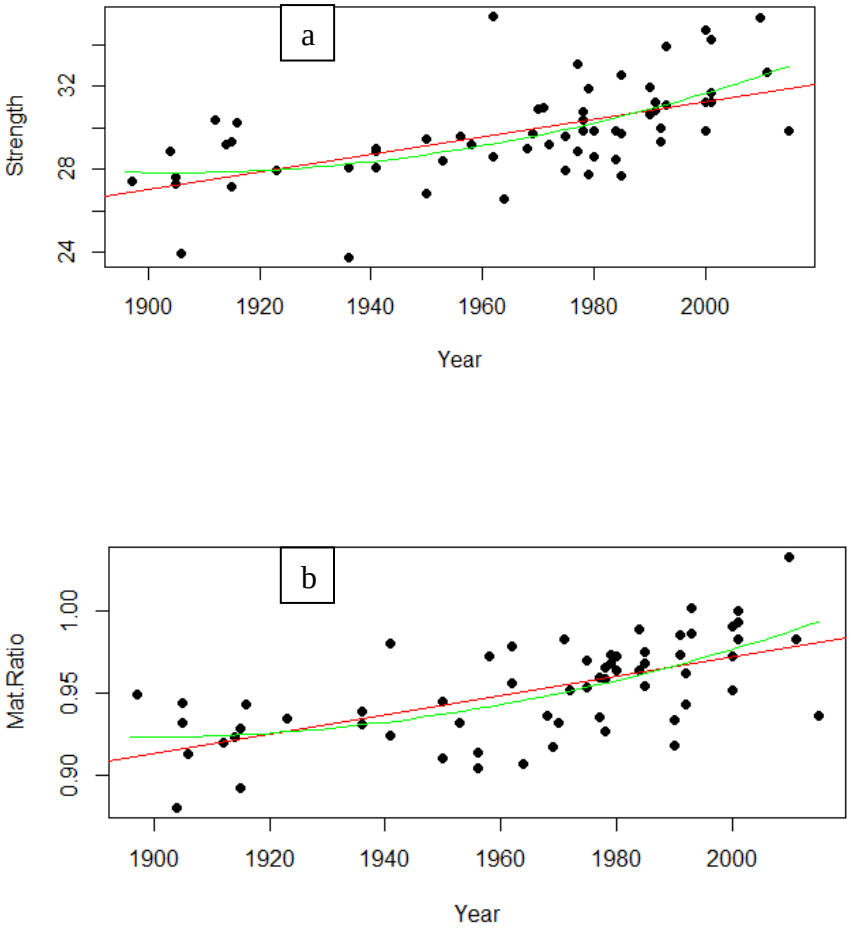


Figure 3.4. Genetic gains for traits that exhibited a better fit with a 2-degree polynomial regression line. a) Genetic gains of HVI fiber strength are plotted against year. b) Genetic gains of AFIS fiber Maturity Ratio plotted against year. c) Genetic gains of AFIS fiber IFC plotted against year. In all plots the red line is the fitted linear regression line, and the green line is the fitted polynomial regression line with 2 degrees. The polynomial regression line is a better fit according to the AIC score for all plots.

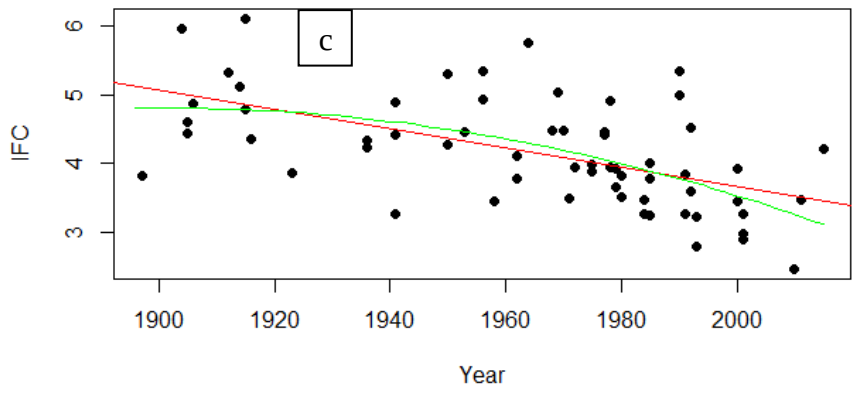


Figure 3.4. Continued.

3.5. References

Arnold, R.W. 1975. The history of adaptation of cotton to the high plains of Texas, 1890-1974. Texas Tech University, Lubbock, TX.

Bates, D., M. Mächler, B. Bolker, and S. Walker. 2014. Fitting linear mixed-effects models using lme4. stat.CO, Ithaca, NY.

Beasley, J.O. 1940. The origin of american tetraploid *Gossypium* species. Am. Nat. 74:285-286.

Benzina, H., E. Hequet, N. Abidi, J. Gannaway, J.-. Drean, and O. Harzallah. 2007. Using fiber elongation to improve genetic screening in cotton breeding programs. Text. Res. J. 77:770-778.

Bourland, F.M., and C.W. Smith. 2001. Registration of Arkot A306 and Arkot A314 germplasm lines of cotton. Crop Sci. 41:2008-2009.

Bowman, D. 2006. Pedigrees of upland and pima cotton cultivars released between 1970 and 2005. Miss Agric For Exp Stn Bull 1155

Bowman, D.T., and O.A. Gutiérrez. 2003. Sources of fiber strength in the US upland cotton crop from 1980 to 2000. J.Cotton Sci 7:164-169.

Brown, H.B. 1936. Cotton varieties recognized as standard commercial varieties. Journal of the American Society of Agronomy 28:69-79.

Brown, H.B. 1927. Cotton: History, species, varieties, morphology, breeding, culture, diseases, marketing, and uses. 1st ed. by harry bates brown .. New York etc. McGraw-Hill book company, Inc., 1927; 1st ed, New York.

Campbell, B.T., S. Saha, R. Percy, J. Frelichowski, J.N. Jenkins, W. Park, C.D. Mayee, V. Gotmare, D. Dessauw, M. Giband, X. Du, Y. Jia, G. Constable, S. Dillon, I.Y. Abdurakhmonov, A. Abdukarimov, S.M. Rizaeva, A. Adullaev, P.A.V. Barroso, J.G. Padua, L.V. Hoffmann, and L. Podolnaya. 2010. Status of the global cotton germplasm resources. Crop Sci. 50:1161-1179.

Cantrell, R.G., C.L. Roberts, and C. Waddell. 2000. Registration of 'Acala 1517-99' cotton. Crop Sci. 40:1200-1201.

Culp, T.W., and D.C. Harrell. 1980. Registration of extra-long staple cotton germplasm (reg. no. GP 150 to GP 154). Crop Sci. 20:291.

Culp, T.W., R.F. Moore, and J.B. Pitner. 1985a. Registration of Pd-1 cotton. Crop Sci. 25:198.

Culp, T.W., R.F. Moore, and J.B. Pitner. 1985b. Registration of Pd-2 cotton. Crop Sci. 25:198-199.

Duncan, E.N., and J.B. Pate. 1964. Registration of Pope Cotton1 (reg. no. 43). Crop Sci. 4:445.

Ewing, E.C. 1965. Deltapine 15, Deltapine staple, Deltapine Smooth Leaf, Fox 4, and Deltapine 45 Cottons (reg. nos. 46, 47, 48, 49, and 50). Crop Sci. 5:199-200.

Green, C.C., and T.W. Culp. 1990. Simultaneous improvement of yield, fiber quality, and yarn strength in upland cotton. Crop Sci. 30:66-69.

Harrell, D.C., and T.W. Culp. 1979. Registration of Pee-Dee-0259 and Pee-Dee-2165 germplasm lines of cotton. Crop Sci. 19:418.

Hequet, E.F., B. Wyatt, N. Abidi, and D.P. Thibodeaux. 2006. Creation of a set of reference material for cotton fiber maturity measurements. Text. Res. J. 76:576-586.

Hertel, K.L. 1953. The stelometer, it measures fiber strength and elongation. Textile World 103:97-260.

Hinze, L.L., D.D. Fang, M.A. Gore, B.E. Scheffler, J.Z. Yu, J. Frelichowski, and R.G. Percy. 2015. Molecular characterization of the *Gossypium* diversity reference set of the US national cotton germplasm collection. Theor. Appl. Genet. 128:313-327.

Hulse-Kemp, A.M., J. Lemm, J. Plieske, H. Ashrafi, R. Buyyarapu, D.D. Fang, J. Frelichowski, M. Giband, S. Hague, L.L. Hinze, K.J. Kochan, P.K. Riggs, J.A. Scheffler, J.A. Udall, M. Ulloa, S.S. Wang, Q. Zhu, S.K. Bag, A. Bhardwaj, J.J. Burke, R.L. Byers, M. Claverie, M.A. Gore, D.B. Harker, M.S. Islam, J.N. Jenkins, D.C. Jones, J. Lacape, D.J. Llewellyn, R.G. Percy, A.E. Pepper, J.A. Poland, K.M. Rai, S.V. Sawant, S.K. Singh, A. Spriggs, J.M. Taylor, F. Wang, S.M. Yourstone, X. Zheng, C.T. Lawley, M.W. Ganal, A. Van Deynze, I.W. Wilson, and D.M. Stelly. 2015.

Development of a 63K SNP array for cotton and high-density mapping of intraspecific and interspecific populations of *Gossypium spp*. G3-Genes Genomes Genetics 5:1187-1209.

Iqbal, M.J., O. 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. Theor. Appl. Genet. 103:547-554.

JASSO, A.H., and L.P. SOLIS. 1994. Registration of Ciano Cocorim-92 cotton. Crop Sci. 34:536.

Jones, K.R. 1998. Breeding history of deltapine 16 and deltapine 50. p. 536-537. *In* Breeding history of Deltapine 16 and Deltapine 50. Beltwide cotton conference, San Deigo, CA. January 5-9, 1998 1998. National Cotton Council, Memphis, TN.

Krowicki, R.S., and H.H. Ramey. 1984. An examination of the digital fibrograph length uniformity index. Crop Sci. 24:378-381.

Lord, E. 1956. Air flow through plugs of textile fibres: Part II. the micronaire test for cotton. Journal of the Textile Institute Transactions 47:T47.

Lubbers, E.L., and P.W. Chee. 2009. The worldwide gene pool of *G. hirsutum* and its improvement. p. 23-52. *In* A. Patterson (ed.) Genetics and genomics of cotton. Springer, New York.

May, O.L., and R.A. Taylor. 1998. Breeding cottons with higher yarn tenacity. Text. Res. J. 68:302-307.

McPherson, R., E. Lubbers, F. Bordelon and J. Schwer. 2000. Phytogen PSC 355 and PSC 952: Conventional, early maturing picker varieties. 2000 Beltwide Conferences, San Antonio, TX. January 4-8, 2000 2000. National Cotton Council, Mephis, TN.

Meredith, W.R. 1993. Registration of Md51ne cotton. Crop Sci. 33:1415.

Miller, P.A., and J.O. Rawlings. 1967. Selection for increased lint yield and correlated responses in upland cotton *Gossypium hirsutum L.* Crop Sci. 7:&.

O'Kelly, J.F. 1950. Registration of improved cotton varieties, IV. Agron. J. 42:53.

R core team. 2016. Vienna: R foundation for statistical computing.

Ramey, H.H. 1966. Historical review of cotton variety development. p. 310-326. *In* Historical review of cotton variety development. 18th cotton improvement conference, Memphis, TN. January 11-12, 1966 1966. National Cotton Council, Mephis, TN.

Sappenfield, W.P., T. Kerr, and W.M. Bugbee. 1972. Registration of delcot 277 Cotton1 (reg. no. 55). Crop Sci. 12:126-127.

Smith, C.W., S. Hague, and D. Jones. 2011. Registration of 'tamcot 73' upland cotton cultivar. Journal of Plant Registrations 5:273-278.

Smith, C.W., R.G. Cantrell, H.S. Moser and S.R. Oakley. 1999. History of cultivar development in the united states. p. 99-173. *In* C.W. Smith, and J.T. Cothren (eds.) Cotton: Origin, history, technology, and production. John Wiley & Sons, New York.

Tyagi, P., M.A. Gore, D.T. Bowman, B.T. Campbell, J.A. Udall, and V. Kuraparthy. 2014. Genetic diversity and population structure in the US upland cotton (*Gossypium hirsutum L.*). Theor. Appl. Genet. 127:283-295.

Webb, R.W. 1932. The Suter-Webb cotton fiber duplex sorter and the resulting method of length-variability measurements. Am.Soc.Test.Mater.Proc. 32:764-771.

Zhang, M., Y.H. Wu, M.K. Lee, Y.H. Liu, Y Rong, T.S. Santos, C. Wu, F. Xie, R.L. Nelson, H.B. Zhang. 2010. Numbers of genes in the NBS and RLK families vary by more than four-fold within a plant species and are regulated by multiple factors. Nucleic Acids Res. 19:6513-6525

4. STABILITY OF HVI AND AFIS TRAITS IN UPLAND COTTON

4.1. Introduction

Understanding the stability of traits in target environments is essential for breeders success. Understanding of stability allows for breeders to make decisions such as how many years, locations, and replications are sufficient to distinguish between superior cultivars, and identify optimal testing environments. Allen et al. (1978) argued that correlations between environments for a trait is important in calculating potential gains in crop improvement, as gains are subject to the environment in which they are tested. Essentially, understanding stability will determine the applicability of gains across environments.

Gossypium hirsutum, or upland cotton, provides 95% of the global cotton fiber production. The US is the largest exporter of upland cotton, and cotton is the number one export commodity of Texas, contributing 1.6 billion dollars to the Texas economy (Texasagriculture.gov). As spinning technologies develop and improve, there is increased demand for superior fiber quality; thus, genetic improvements of cotton fiber quality is becoming increasingly important. In a survey of cotton producers, an important priority listed that producers want from science is the improvement of fiber quality (pers. comm. Kater Hake, Cotton Inc., 2016). Determination of fiber quality is performed generally by High Volume Instrument (HVI), which is fairly cheap and measures the properties of a bundle of fibers (See Chapter 1). The classification system of US cotton has been described by Cotton Inc (www.cottoninc.com/fiber/quality/US-Fiber-Chart/Ratings-Of-Fiber-Properties/) (Table 4.1). Advanced Fiber Information System (AFIS) technologies are more

recent in scope and provide analysis of the properties of individual fibers rather than a bundle of fibers as per HVI, and can be used to separate the confounding variables of the HVI trait micronaire, by measuring both maturity and fineness of fibers (Hequet et al., 2006). A problem with acquiring AFIS data is cost. AFIS can cost as much as six times more per sample than HVI (https://www.depts.ttu.edu/pss/fbri/fee09.pdf). Understanding the stability of AFIS and HVI data will help breeders optimize program resources in utilizing this technology.

There is a large genetic component controlling fiber quality traits, both HVI and AFIS traits, as indicated by high heritabilities and is well documented. Dabbert et al (2017) looked heritabilities across varying temperatures and varying water deficits in Georgia, Texas, and Arizona, reporting that heritabilities between environments did not change substantially across environments for HVI fiber quality traits. Other studies measured heritabilities in similar locations, but had larger variation in heritability than the Dabbert et al. (2017) study. All studies still reported high heritabilities even though they were conducted in different states (Braden and Smith, 2004; Hugie et al., 2017; Zeng and Bechere, 2017). It is possible the larger variation in heritabilities is due to crosses more than environment. The purpose of this study is to look more into the GxE effect, stability, and the consequences of selecting top performing cultivars in one environment over another environment using HVI and AFIS determined fiber quality properties.

4.2. Materials and Methods

The population used in this study contained 117 genotypes, consisting of 13 released cultivars from Texas A&M's Cotton Improvement Lab, 71 previously released obsolete US

improved cultivars, and 33 RILs from 5 different intraspecific crosses that were selected bi-directionally for high quality and low quality using HVI upper half mean length (UHML) and strength parameters. This population was planted in a randomized complete block design (RCBD) in 2016 and 2017 in Weslaco, TX at the Texas A&M AgriLife Research and Extension Center and in Corpus Christi, TX at the Texas A&M AgriLife Research and Extension Center. In 2016, three replications were used at both locations, and in 2017 two replications were used at both locations. Soil type at Weslaco is a Hidalgo sandy clay loam, a fine-loamy, mixed, active, hyperthermic Typic Calciustolls, and at Corpus Christi soil type is a Houston black clay, a fine smectitic, thermic Udic Haplustert. Normal cotton production practices were used in all trails, with furrow irrigation used in Weslaco, TX. No irrigation was used at the Corpus Christi testing location.

Boll samples were randomly harvested from plots with 30 bolls hand harvested from the first fruiting limb position in the middle of the fruiting zone. Fiber was ginned using 8-saw laboratory gins, with each replication ginned on a single gin. Fiber samples were sent to the Fiber and Biopolymer Research Institute at Lubbock, TX where phenotyping was performed using HVI and AFIS. The traits used in this study from HVI were upper half mean length (UHML), fiber bundle strength (strength), elongation, and uniformity. AFIS traits evaluated were standard fineness, maturity ratio (maturity), and short fiber content by number (SFC). Definition of these traits are given in chapter 2.

An Analysis of Variance (ANOVA) was calculated for each trait using the lm function in R (R core team, 2016), and Fisher's LSD was used for multiple comparisons of means of cultivars for the combined analysis and each individual environment using the

function LSD.test from the Agricolae package in R (De Mendiburu, 2014). The top 20% from the combined analysis and each individual analysis was selected numerically from the lsmeans. These selections were compared between each analysis to determine if the selected cultivars overlapped between analyses. This comparison was accomplished in two ways: 1) which cultivars selected in one analysis were also selected in other analyses 2) which cultivars selected in one analysis were not in the top 20% selected numerically, but where within the Fisher's LSD rank of the cultivars selected in the top 20% of the other analyses; therefore, making the differences insignificant. Spearman's correlation was used to determine the strength of ranking of cultivars across each environment.

Stability regression was calculated according to Eberhart and Russell (1966) using R. In summary, an environmental index was calculated by subtracting the mean of each location from the grand mean. The environmental index is then regressed to the genotype mean for each environment using the following formula: $Y_{ij} = \mu_i + \beta_i I_j + \delta_{ij}$, Where Y_{ij} is the cultivar mean of the i^{th} cultivar at the j^{th} environment, μ_i is the mean of the i^{th} cultivar over all environments, β_i is the regression coefficient that measures the response of the i^{th} cultivar to varying environments (referred to the stability coefficient in this paper), I_j is the environmental index at the j^{th} environment, and δ_{ij} is the deviation from the regression of the i^{th} cultivar at the j^{th} environment. Cultivars were then classified into categories of fiber ratings as determined by Cotton Incorporated's classifications (www.cottoninc.com/fiber/quality/US-Fiber-Chart/Ratings-Of-Fiber-Properties/) (Table 4.1) with the exception of SFC. SFC values were separated into quartiles with the lowest SFC being in the 1^{st} quartile and the highest SFC being in the 4^{th} quartile. Standard fineness was

used in replace of fineness and is calculated as AFIS fineness divided by AFIS maturity ratio. The same rankings determined by Cotton Incorporated for fineness were used for classifying standard fineness. Stability coefficients for each cultivar were grouped according to classification for each trait, and ANOVA was used to determine if there was a significant difference in stability based on trait values. This stability analysis was applied to all traits with the exception of maturity, because 112 cultivars in this study were classified as mature and only 5 were classified to another group which was very mature. This did not allow for a fair comparison of multiple classes. The lack of variation in maturity is due to boll sampling method. Boll samples are taken from the first fruiting limb position in the middle the fruiting zone on each plant which insures a uniform maturity.

4.3. Results and Discussion

HVI traits UHML, uniformity, and strength of these 117 genotypes were significantly affected by environment, genotype and GxE at p value < 0.001 in the combined analysis (Table 4.2). Elongation showed significant variation across genotypes (p value < 0.001), but did not show a significant variation for environment or the interaction of genotype by environment (GxE). AFIS traits maturity, SFC, and standard fineness showed significant variation across genotypes with a p value < 0.001, and a significant GxE variation for maturity with a p value < 0.001, and SFC and standard fineness with a p value < 0.001. Maturity, standard fineness, and SFC significantly varied across environments (p value < 0.001 for maturity and standard fineness, and p value < 0.05 for SFC).

The greatest contribution to total sum of squares (TSS) for all traits was genotype (Table 4.3). UHML had the largest genotypic contribution to TSS with genotypes

accounting for 80.64% of TSS. This was seven times greater than the second largest contributor to TSS for UHML, which was environment at 11.37%. The lowest genotypic contribution to TSS was for SFC at 42.33%. SFC also had the highest contribution of environment at 33.39. The largest contribution to TSS for GxE was with elongation at 16.02%; however as mentioned above this was the only trait that did not have a significant effect for GxE. This is due to the higher residual error. Elongation also had the highest contribution for residual error to TSS of any traits at 33.48%. For all traits the largest source of variation comes from genotype, with environment and GxE contributing much less. Campbell and Jones (2005) reported similar results only for strength and elongation involving eight commercial cultivars in performance trails, with environment being the greatest source of variation for length and uniformity. Campbell et al. (2012) found that environment was the largest source of variation for all HVI traits. Both of these studies evaluated cultivars in more environments with 12 and 14 different environments evaluated respectively, which could lead to more variation for environmental effect. Another reason for the differences could be the populations evaluated. Campbell and Jones (2005), described eight elite commercial cultivars, which likely were comparable for fiber properties. Campbell et al. (2012) evaluated 82 released cultivars that represented the history of the Pee Dee cotton germplasm enhancement program. In evaluation of those same cultivars for genetic gains Campbell et al. (2011) found that fiber properties were a priority for selection in the beginning of the Pee Dee program, but switched priorities to yield and maintaining fiber quality. By switching priorities to maintaining fiber quality, the variation of fiber quality was likely limited in later released cultivars. In the population used in this

study, cultivars ranged from over a hundred years of breeding efforts in the US, with cultivars and germplasm lines developed through Texas A&M's Cotton Improvement Lab resulting with a high selection priority on fiber improvement. The combination of historical cultivars with Texas A&Ms genotypes created a population with considerable variation for fiber quality traits among cultivars, which can be seen in the histograms provided in Figure 4.1.

The results from the ANOVA of stability coefficients calculated according to Eberhart and Russel (1966) for genotypes assigned to different classes according to Table 4.1 is provided in Table 4.4. The only traits with a significant difference in stability coefficients among the classes was for standard fineness and SFC. As indicated in Table 4.1, 37 genotypes ranked as Fine, 75 genotypes ranked as Average, and only 5 genotypes are ranked as Coarse. To ensure that the significant differences were not the result of a sampling bias, as the Coarse class only had 5 genotypes, a Welch Two Sample t-test was used to compare stability coefficients only between Fine and Average class genotypes. The Welch Two-Sample t test indicated that there those stability coefficients were different (p value < 0.01). The mean stability coefficient for the Fine class was .74 and the mean stability coefficient for the Coarse class was 1.11, suggesting that finer fibers are more stable than coarser fibers.

For SFC, Fisher's LSD was used to compare the classes to determine which classes were more stable than the others. There was a significant difference between very low SFC and low SFC with high SFC. The mean stability coefficients for the very low SFC and high SFC classes were .86, .97, and 1.16 respectively. The lower the stability coefficient is

interpreted as more stable as this means there is less difference in genotypes across environments. This indicates that genotypes with lower SFC are more stable than those with higher SFC. For both SFC and standard fineness, the tendencies for stability is favorable for the breeder. As improvements are made for both traits, improvements for stability should occur concomitantly. Campbell and Jones (2005) reported a similar trend for strength, where lower-strength genotypes were less stable than higher-strength genotypes; however, this was not observed in this study. Geng et al. (1987), used a quality score to compare stability coefficients and found that as the quality score increased so too did the stability. The quality score used length parameters, uniformity, and strength in the calculation. Again, all comparable traits in this study to the traits used to calculate the quality score in Geng et al. (1987) did not show more stability as quality improved.

Spearman's correlation coefficients were used to assess strength of ranking among genotypes across environments and are shown in Table 4.5. Correlations generally were high for all traits. The lowest correlations existed with elongation and SFC, with the lowest correlation being between Corpus Christi 2016 and Corpus Christi 2017 for elongation at 0.51. The highest correlations were for UHML, strength, and standard fineness with an average correlation of 0.89, 0.90, and 0.90, respectively, with little difference between the highest correlation and the lowest correlations with differences of 0.06, 0.06, and 0.08, respectively. This shows that ranking doesn't appear to change drastically between any environment for these traits. The lower correlations for elongation and SFC could be contributed to the lower proportion of variation attributed to genotype observed in Table 4.3. To further clarify, the top 20% of genotypes where selected numerically for each trait at

each environment and from a combined analysis across environments. The list of selected genotypes from each location were compared to each other to see which genotypes were on the same list. Fisher's LSD was also calculated for each location and the combined analysis to check if genotypes not in the top 20% numerically on each list where still within the LSD grouping from the selected genotypes, which indicate that they are not different. UHML was by far the most consistent with selection within the environments (Figure 4.2). The top 20% selected numerically consist of a selection of 23 genotypes, and 21 of the selected genotypes were shared between all locations for UHML. Only one genotype was in the top 20% numerically at one location and not within the LSD of the top 20% from another location. For all fiber traits, there were few genotypes that were inconsistent across trials. The trait with the most inconsistency of genotypes that were not within the LSD of the top 20% in one environment, but were within the top 20% numerically of another environment was standard fineness with seven genotypes; however, this is still relatively few, and the inconsistency of the genotypes appears random. Maturity had the least overlap of top 20% numerically selected genotypes. Only 6 of the 23 genotypes where in the top 20% numerically between all environments; however, there were only two genotypes that were inconsistent by being out of the LSD with another environment, and for each of these genotypes was only inconsistent with one other environment. These data indicate that selection for these fiber quality traits could be performed in any of these environments with similar ability to obtain the highest quality genotypes, and that these fiber quality traits are highly stable across environments. The conclusion of Campbell et al. (2012) that the

significance in GxE for fiber quality is a product of magnitude and not rank change, which was substantiated by Ng et al. (2013) was further affirmed by the findings in this study.

4.4. Conclusion

This study shows that fiber traits are highly stable in the environments tested. It was shown that selection for fiber quality traits in any of these environments would result in a comparable list of selected genotypes. This reaffirms the findings of both Campbell et al. (2012) and Ng et al. (2013) that discuss that GxE for fiber quality traits is more significant for magnitude rather than rank changes. This study also shows no differences in stability as fiber trait values change with the exceptions of AFIS's standard fineness and SFC. For these traits the changes in stability for these traits favors the breeders as stability increases as fiber quality trait values improves. The lack of change in stability for the other fiber quality traits counters the opposing conclusions of Geng et al. (1987) and Campbell and Jones (2005). This study concludes that selection for these fiber quality traits can be adequality performed in just one year in either of these locations. This is beneficial to the breeder in reducing the necessary number resources used while still attaining accurate selections.

Table 4.1. Ratings of fiber traits and grouping of populations by ratings established by Cotton Incorporated (www.cottoninc.com/fiber/quality/US-Fiber-Chart/Ratings-Of-Fiber-Properties/), with the exception of SFC, which is ranked and grouped by quartiles.

Range	Classification	# of Cultivars in each Class
Upper Half Mean Length (in)		
Bellow 0.99	Short	3
0.99-1.10	Medium	36
1.11-1.26	Long	54
Above 1.26	Extra Long	24
Fiber Elongation (%)		
Below 5.0	Very Low	0
5.0-5.8	Low	5
5.9-6.7	Average	59
6.8-7.6	High	37
Above 7.6	Very High	16
Uniformity (%)		
Below 77	Very Low	0
77-79	Low	3
80-82	Average	51
83-85	High	46
Above 85	Very High	17
Standard Fineness (Unit)		
Below 135	Very Fine	0
135-175	Fine	37
175-200	Average	75
200-230	Coarse	5
Above 230	Very Coarse	0
Fiber Maturity Ratio (%)		
Below 0.7	Uncommon	0
0.7-0.8	Immature	0
0.8-1.0	Mature	112
Above 1.0	Very Mature	5
Fiber Strength (grams/tex)		
23 and below	Weak	0
24-25	Intermediate	3
26-28	Average	31
29-30	Strong	33
31 and above	Very Strong	50
Short Fiber Content (%)		
1st Quartile	Very Low	29
2nd Quartile	Low	29
3rd Quartile	Medium	29
4th Quartile	High	40

Table 4.2. Mean Squares from ANOVA of combined analysis of the 117 gentoypes for fiber quality traits.

		Mean Squares						
Source	DF	UHML	Elongation	Uniformity	Std. Fine	Strength	Maturity	SFC
Env	3	0.68686**	54.092	196.371*	41.98**	260.024**	0.075671**	2900.83*
Block	2	0.00564**	3.361*	5.009**	0.36	0.827	0.000757	104.99***
Genotype	116	0.12596***	6.645***	23.821***	1350.5***	112.827***	0.01067***	95.39***
GxE	348	0.00178***	0.832	1.063***	29.28***	3.155***	0.000477**	7.72***
Residuals	676	0.00121	0.895	0.789	17.02	1.932	0.00038	5.14

*Significance at 0.05
**Significance at 0.01
***Significance at 0.001

Table 4.3. Proportion of total sum of squares from the combined ANOVAs for the seven fiber traits used in study. Significance level is given from the F-test calculated in the ANOVAs.

	Proportion of Sum of Squares						
Source	UHML	Elongation	Uniformity	Std. Fine	Strength	Maturity	SFC
Env	11.37**	7.49	13.81*	0.07**	4.79**	12.02**	33.29*
Block	0.06**	0.37*	0.23**	0.00	0.01	0.08	0.80***
Genotype	80.64***	42.65***	64.78***	87.77***	80.42***	65.53***	42.33***
GxE	3.41***	16.02	8.67***	5.71***	6.75***	8.78**	10.28***
Residuals	4.51	33.48	12.50	6.45	8.03	13.59	13.29

*Significance at 0.05
**Significance at 0.01
***Significance at 0.001

Table 4.4. Mean squares and significance levels from ANOVAs calculated to determine differences in Eberhart and Russel (1966) regression coefficient between the different fiber ranking classes for different fiber traits.

	Mean Squares					
Source	UHML	Elongation	Uniformity	Standard Fineness	Strength	SFC
Classes	0.123299	0.21526	0.37715	1.7845**	0.86688	0.4315*
Residuals	0.092153	0.86058	0.25475	0.3076	0.48561	0.12053

*Significance at 0.05
**Significance at 0.01

Table 4.5. Spearman Correlations for traits between each environment. CC is abbreviation for Corpus Christi and W is the abbreviation for Weslaco. The two digit number represents the year at the respective location.

UHML				
	CC16	W16	CC17	W17
CC16	1.00			
W16	0.92	1.00		
CC17	0.87	0.88	1.00	
W17	0.90	0.90	0.86	1.00
Elongation				
	CC16	W16	CC17	W17
CC16	1.00			
W16	0.67	1.00		
CC17	0.51	0.55	1.00	
W17	0.55	0.70	0.63	1.00
Uniformity				
	CC16	W16	CC17	W17
CC16	1.00			
W16	0.86	1.00		
CC17	0.76	0.75	1.00	
W17	0.80	0.83	0.72	1.00
Standard Fineness				
	CC16	W16	CC17	W17
CC16	1.00			
W16	0.94	1.00		
CC17	0.87	0.89	1.00	
W17	0.89	0.92	0.91	1.00
Strength				
	CC16	W16	CC17	W17
CC16	1.00			
W16	0.92	1.00		
CC17	0.88	0.86	1.00	
W17	0.90	0.91	0.91	1.00
Maturity				
	CC16	W16	CC17	W17
CC16	1.00			
W16	0.88	1.00		
CC17	0.78	0.80	1.00	
W17	0.85	0.88	0.81	1.00
SFC				
	CC16	W16	CC17	W17
CC16	1.00			
W16	0.78	1.00		

CC17	0.66	0.65	1.00	
W17	0.76	0.70	0.67	1.00

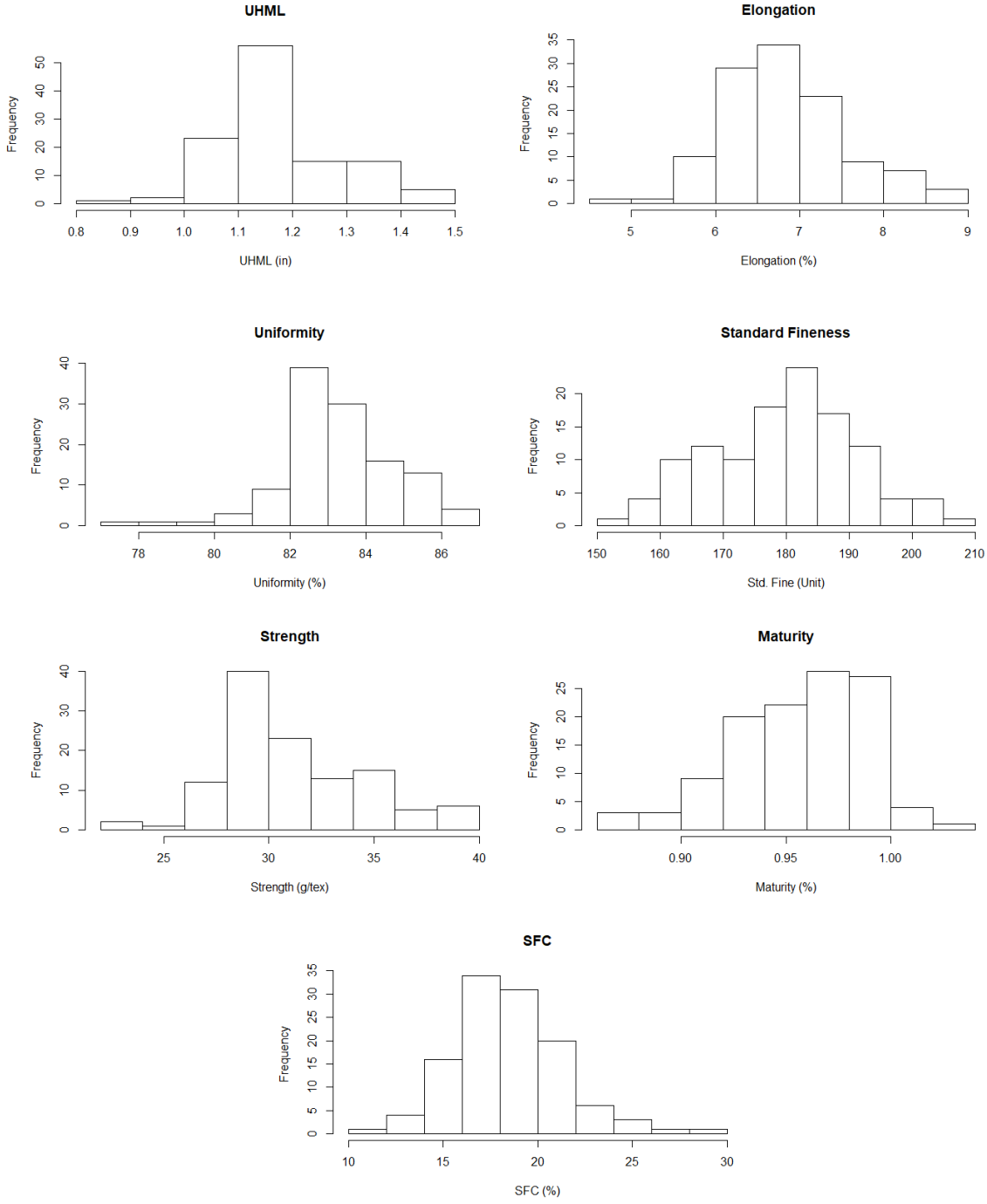
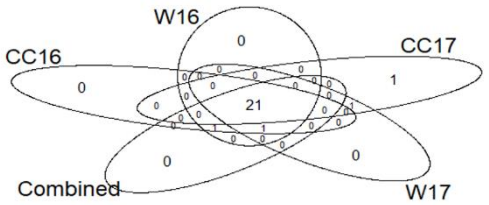


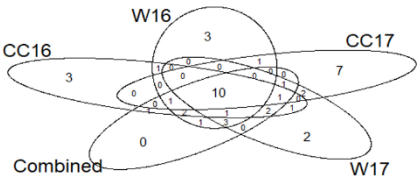
Figure 4.1. Histograms of traits for population used in this study.

HVI UHML



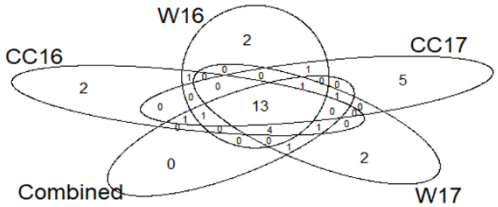
Cultivar	Combined	CC16	CC17	W16	W17
KH-13-021-01			O	X	

HVI Elongation



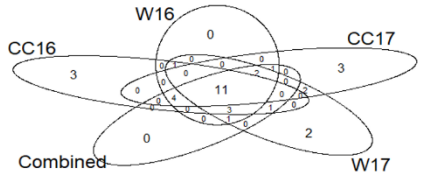
Cultivar	Combined	CC16	CC17	W16	W17
KH-13-267-04		X	O	X	
KH-13-047-03			O	X	
Coker's Clevewilt 3			X	O	
DES 119		O			X

HVI Uniformity



Cultivar	Combined	CC16	CC17	W16	W17
MD51ne	X			O	X
Acala Maxxa		X	O		
Lightning Express		X	O		
Coker 310			O	X	

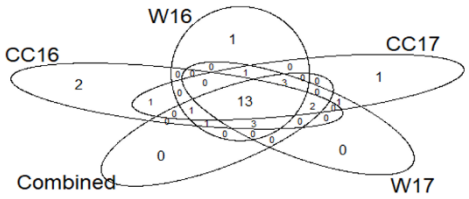
AFIS Standard Fineness



Cultivar	Combined	CC16	CC17	W16	W17
Des 119	O	O	O	O	XXXX
LA887	X	O		X	X
Durango		O	XXX	O	O
Westburn		XX	O		O
Big Boll Triump			O	X	
Meade Clean Seed	X		O		
Blight Master				X	O

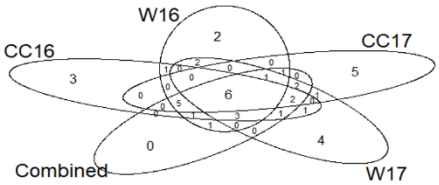
Figure 4.2. Venn Diagrams showing number of cultivars shared between location when selected by top 20% numerically for each trait labeled in the figure. The table below the diagram shows cultivars that were in the top 20% numerically at one environment, but were not within Fisher's LSD of the top 20% at another environment. The environment where the Cultivar was cultivar was within the top 20% is indicated by O, and the environment where the cultivar was not within LSD of the top 20% is indicated by X.

HVI Strength



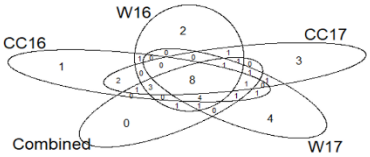
Cultivar	Combined	CC16	CC17	W16	W17
KH-13-118-03		O	O	XX	
KH-13-142-01		O	O	XX	
TM-1	X		X	O	
KH-13-122-04			O	X	

AFIS Maturity



Cultivar	Combined	CC16	CC17	W16	W17
PSC 355				O	X
DES 119			O	X	

AFIS SFC



Cultivar	Combined	CC16	CC17	W16	W17
KH-13-142-01	X	X		O	
TAMB182-33	X			O	X
Del Cerro	X	X	O		
TX 2094	X		O		

Figure 4.2. Continued.

4.5. References

Braden, C.A., and C.W. Smith. 2004. Fiber length development in near-long staple upland cotton. Crop Sci. 44:1553-1559.

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

Campbell, B.T., P.W. Chee, E. Lubbers, D.T. Bowman, W.R. Meredith Jr., J. Johnson, and D.E. Fraser. 2011. Genetic improvement of the Pee Dee cotton germplasm collection following seventy years of plant breeding. Crop Sci. 51:955-968.

Campbell, B.T., P.W. Chee, E. Lubbers, D.T. Bowman, W.R. Meredith Jr., J. Johnson, D. Fraser, W. Bridges, and D.C. Jones. 2012. Dissecting genotype x environment interactions and trait correlations present in the pee dee cotton germplasm collection following seventy years of plant breeding. Crop Sci. 52:690-699.

Dabbert, T.A., D. Pauli, R. Sheetz, and M.A. Gore. 2017. Influences of the combination of high temperature and water deficit on the heritabilities and correlations of agronomic and fiber quality traits in upland cotton. Euphytica 213:6.

De Mendiburu, F. 2014. Agricolae: Statistical procedures for agricultural research. 1.

Eberhart, S.T., and W.A. Russell. 1966. Stability parameters for comparing varieties. Crop Sci. 6:36-40.

Geng, S., Q.F. Zhang, and D.M. Bassett. 1987. Stability in yield and fiber quality of California cotton. Crop Sci. 27:1004-1010.

Hequet, E.F., B. Wyatt, N. Abidi, and D.P. Thibodeaux. 2006. Creation of a set of reference material for cotton fiber maturity measurements. Text. Res. J. 76:576-586.

Hugie, K.L., C.W. Smith, K.S. Joy, and D.C. Jones. 2017. Divergent selection for fiber length and bundle strength and correlated responses in cotton. Crop Sci. 57:99-107.

Ng, E., K. Jernigan, W. Smith, E. Hequet, J. Dever, S. Hague, and A.M.H. Ibrahim. 2013. Stability analysis of upland cotton in Texas. Crop Sci. 53:1347-1355.

R core team. 2016. Vienna: R foundation for statistical computing.

Zeng, L., and E. Bechere. 2017. Correlated selection responses of fiber properties measured by high volume instrument and advanced fiber information system in upland cotton. Euphytica 213:278.

5. GENOMIC PREDICTION IN UPLAND COTTON

5.1. Introduction

As genomic technologies advance, Plant breeders are provided with new tools to facilitate selection of favorable alleles. Genomic prediction is one such tool that utilizes molecular markers to detect genetic variation for the prediction of phenotypic performance. It is different from more traditional QTL analysis in that it is a multi-variate method, and allows for evaluation of multiple marker effects simultaneously (Meuwissen et al., 2001). This technique has application when genotyping is more affordable than phenotyping, and more recently has showed promise in identifying favorable alleles in germplasm collections (Yu et al., 2016; Thorwarth et al., 2018).

Techniques for measuring cotton fiber quality have evolved throughout the twentieth century and continues today. Early breeding work focused on field evaluation and consisted of subjectively evaluating cotton fibers for length and strength by human observation. Standardization came in the US Cotton Futures Act of 1918 for staple length, but this also only relied on subjective measurements. Objective measurement for strength was developed in 1926, length and length distribution in 1932, and bundle fiber strength and elongation in 1953, but these techniques were stand-alone measurements and were slow and tedious (See Chapter 1). High Volume Instrumentation (HVI) combined many fiber quality measurements on a single machine. Developed in 1968, HVI wasn't widely implemented into either private or public breeding programs until the late 1980s. HVI analyzes a bundle of fibers of a given weight that is taken from each sample, and has evolved over the years for

determination of upper half mean length (UHML), length uniformity, strength, elongation, and micronaire relatively rapidly (See Chapter 2). Modern breeding programs utilize HVI at all stages of the breeding process, and it is currently the industry standard. AFIS was developed in the late 1980s to evaluate fiber quality on a single fiber basis (Bragg and Shofner, 1993). This system can evaluate fiber length, length distribution, fineness, maturity, and neps (Williams and Yankey, 1996); however, it is more time consuming and higher cost than HVI. The higher cost and lack of general acceptance within the community prevents the wide use of this phenotyping method.

HVI and AFIS are both used to predict fiber spinning quality and yarn properties (Faulkner et al. 2012), which is the goal of improving fiber quality. Evaluation of spinning quality and yarn properties requires a large sample of fibers, is time consuming, and expensive. A mini-spin protocol developed by Hequet and outlined by Joy et al. (2010), requires a smaller sample, but is still large compared with HVI and AFIS, time consuming, and expensive, which prevents it from being implemented at early stages in a breeding program.

There has been more than 100 years of breeding efforts in the creation of the US improved cotton germplasm. Although fiber phenotyping techniques have been implemented during that time, early methods were subjective or the slow and tedious nature of data collection prevented analysis. As a result, many sources of potential fiber quality alleles likely have been overlooked. This study accomplishes three objectives: 1) Evaluate the feasibility of genomic prediction in upland cotton for fiber quality traits; 2) Evaluate the application of genomic prediction in obsolete US improved cultivars from the USDA cotton

germplasm collection; and 3) Evaluate the feasibility of using genomic prediction on a selection index developed from HVI and AFIS parameters for yarn work to break.

5.2. Materials and Methods

The population used in 2016 consisted of 128 genotypes that contained 74 previously released obsolete US improved cultivars representing the different cotton growing regions throughout the US, 10 current commercial cultivars, 11 released cultivars and germplasm lines from Texas A&M's Cotton Improvement Lab (CIL), and 33 inbred lines developed by the CIL that were selected bi-directionally for high quality and low quality using HVI upper half mean length (UHML) and strength parameters from five internal biparental populations as described by Hugie et al. (2017) The five internal biparental populations that gave rise to the 33 bidirectional inbred lines were derived from four parents, TAM 03B182-33 (Smith et al. 2009), TAM 06WE-62-04, TAM 04SID842 (an interspecific derived breeding line), and Tamcot 22 (Thaxton and Smith, 2005). TAM 03B182-33, TAM 04SID842, and TAM 06WE-62-4 which contain fiber quality alleles accumulated through decades of pedigree breeding and selection pressure for fiber length and strength, and Tamcot22 (PI 635877) is a released high yielding cultivar, which came out of the same program and contains excellent fiber properties. The population used in 2017 consisted of the same genotypes along with 47 additional released obsolete US improved cultivars representing the different cotton growing regions throughout the US. These respective populations were planted in a randomized complete block design (RCBD) in 2016 and 2017 at the Texas A&M AgriLife Research and Extension Center and in Corpus Christi, TX at the Texas A&M AgriLife Research and Extension Center in Weslaco, TX. In 2016,

128 genotypes in three replications were used at both locations, and in 2017, 174 genotypes in two replications were used at both locations. Soil type at Weslaco is a Hidalgo sandy clay loam, a fine-loamy, mixed, active, hyperthermic Typic Calciustolls, and a Houston black clay, a fine smectitic, thermic Udic Haplustert at Corpus Christi. Normal cotton production practices were used in all trials, with furrow irrigation used in Weslaco, TX. No irrigation was applied at the Corpus Christi testing location.

Boll samples were randomly harvested from plots with 30 bolls hand harvested from the first fruiting limb position in the middle of the fruiting zone. Boll samples were ginned using 8-saw laboratory gins, with each replication ginned by a single gin. Fiber samples were sent to the Fiber and Biopolymer Research Institute at Lubbock, TX where phenotyping was performed using HVI and AFIS. The traits used in this study from HVI were upper half mean length (UHML) and strength, and the traits used from AFIS were length by number L(n), standard fineness, and short fiber content by number (SFC). An index was created from all HVI and AFIS traits to predict yarn work to break.

For yarn quality testing the 128 genotypes from 2016 were also planted that year at the Texas AgriLife Research Farm near College Station, TX on a Westwood silt loam, a fine-silty, mixed, superactive, thermic Udifluventic Haplustepts integrated with Ships silty clay, a very fine, mixed, active, thermic Chromic Hapluderts. Normal cotton production practices, and furrow irrigation was used. This trial was planted in an RCBD with three replications. Plots were harvested with a one-row spindle picker modified for single plot harvest, and the three replications were bulked by genotype. Samples were ginned on a 21-

saw laboratory gin and sent to the Fiber and Biopolymer Research Institute in Lubbock, TX for mini-spinning analyses (Joy et al. 2010).

Tissue samples were collected from young leaves, and DNA was extracted using a modified CTAB (cetytrimethylammonium bromide) method described by Zhang et al. (2010). Genotyping was performed using the Illumina® 63K SNP array (Hulse-Kemp et al., 2015). SNP markers were removed 1) when markers were non-polymorphic, 2) greater than 10 % of SNP calls were missing in population, 3) minor allele frequency was less than .03, and 4) heterozygosity of marker was greater than 10 %. After marker filtering, 20,045 high quality SNPs remained.

For all traits, Empirical Best Linear Unbiased Predictors (EBLUP) were calculated using lme4 package in R (Bates et al., 2015). Genomic prediction was performed in the GAPIT R package (Lipka et al., 2012), where a Genomic Best Linear Unbiased Predictor (GBLUP) model was calculated using a genomic relationship matrix (VanRaden, 2008) and the first three principal coordinates calculated from the marker data used as fixed effects covariates. Genomic prediction models were validated using bootstrap validations, where one-fifth of the population was randomly selected as the test set, and the rest were used as the training set. The bootstraps were run for 1,000 iterations. The mean prediction value for each genotype was then correlated to actual value to determine prediction accuracy.

To evaluate the effects of population structure, fastStructure (Raj et al., 2014) was applied using the default settings and the prior argument set to simple. K-values of 1 to 10 were evaluated for optimal K using the chooseK function. If the optimal K was determined to be 1, the process was repeated for K-values of 1 to 3 with the prior argument set to

logistic, and the chooseK function used again to identify optimal K. The K value indicates how many subpopulations are expected within the full population. This procedure was used on the entire population, and on a subset of the population created for the purpose of inducing population structure. The subset consisted of the 96 obsolete US improved cultivars and 24 of the bidirectional inbreds.

To evaluate the effects of population structure on prediction accuracy, genomic prediction was performed two subset populations. These populations consisted of the obsolete US Improved cultivars and the bidirectional inbreds with one population containing highs and lows from the bidirectional inbreds for UHML, and the other containing highs and lows from the bidirectional inbreds for strength. In both subsets, genomic prediction was performed for the obsolete US improved lines only, then a series of substitution of obsolete US improved lines with the Texas A&M bidirectional inbreds at 5, 10, 15, and 20 percent substitution. Using this substitution method insured that changes in prediction accuracy were not the result of changes in population size, as population size was held constant at each varying percent substitution. This was performed randomly in each bootstrap iteration by randomly choosing an even number of the obsolete US improved lines to be removed that corresponded to the percent of substitution, and randomly choosing a pair of inbreds to replace them. A pair consisted of both a high-quality and a low-quality selection from the same cross. This analysis will be referred to as the inbred substitution analysis in further discussion.

Random forest was used to develop a selection index using HVI and AFIS data to predict yarn work to break with the randomForest package in R (Liaw and Wiener, 2002)

using default arguments. To identify which HVI and AFIS variables were most important the out of bag error rate (OOB) was used. The prediction accuracy and variation of prediction accuracy was determined by using 100 boot strap iterations for each addition of the traits in order of importance as determined by OOB. In each iteration $1/5^{th}$ of the 128 genotypes were selected as the test set, and the remaining genotypes were used to build the model. The best fit model was used to predict yarn work to break for all genotypes in the population, and a genomic prediction analysis was performed on determined trait values using same protocol described earlier. Prediction accuracy was determined by correlation of predicted value with the actual value from the 128 genotypes from which yarn was spun.

5.3. Results and Discussion

The results from the genomic prediction analysis on the full population are shown in Figure 5.1. The correlation for UHML, strength, standard fineness, L(n), and SFC for the full population were 0.73, 0.71, 0.72, 0.63 and 0.48, respectively. However, these values are not consistent when breaking down the germplasm in different groups. Looking at the correlations between predicted value and actual value for the Texas A&M germplasm and the obsolete US improved cultivars separately from the same analysis shows that the prediction accuracies for the group were different than prediction accuracies of the whole. For UHML, strength, and L(n), the predictions accuracies for both groups were less than the prediction accuracies for the whole population. The prediction accuracies for the Texas A&M germplasm was 0.58, 0.54, and 0.60 for UHML, strength, and L(n) respectively, and the prediction accuracies for the obsolete US improved cultivars was 0.56 for both UHML and strength, and 0.54 for L(n). Standard fineness showed similar correlations for both the

whole population and the Texas A&M germplasm at 0.72 and 0.70 respectively. However, the obsolete US improved cultivars showed a correlation of 0.54. UHML, strength, L(n), and standard fineness all show population structure differences detectable by the phenotypic data, with the Texas A&M germplasm having superior fiber quality compared with the obsolete US improved cultivars. This can be seen clearly in Figure 5.1, and a t-test performed on these traits all showed significant differences (p value < 0.001). For the Texas A&M material, 33 of the 44 genotypes were developed from bi-directional selection for high quality and low-quality cotton fiber quality. Again, the parents were developed from decades of pedigree breeding and selection pressure for fiber length and strength, which putatively resulted in the accumulation of alleles for these traits. Even though the 33 inbreds were bi-directionally selected, the CIL genotypes selected for low-quality were generally superior in regards to UHML, strength, L(n), and standard fineness, indicating that alleles for these traits are nested within the population structure. The phenotypic structure causes the inflated correlations for the whole populations as it creates more spread in the data. This is shown in simulated data (Figure 5.2). A simulated data set was created around the formula $Y = \beta_0 + \beta_1 X + error$, where Y is the actual data value, β_0 is the intercept equal to zero, β_1 is the slope equal to one, X is randomly generated numbers split into two groups: the first group was between zero and 0.5, and the second group was between 0.5 and one, and the error is randomly generated numbers from a uniform distribution with a mean of zero and a standard deviation of 0.23. Since the error is random and uniform around zero the predicted values for Y ($\hat{Y}$), were calculated from the same formula, with the error equal to zero. In the simulated data the correlations for the whole data set was 0.73, but the correlations for the groups

within the data set were less at 0.53 and 0.55. This is consistent with what is seen in the empirical fiber data for UHML, strength, L(n), and standard fineness.

SFC had different correlations between the whole population and the two groups, with the whole population showing a correlation of 0.48 and the Texas A&M germplasms and obsolete US cultivars showing 0.73 and 0.24 respectively. Unlike the other traits, there is no clear population structure according to the phenotypic data for SFC in Figure 5.1, and the results of the t-test showed no significant difference (p value = 0.63) between Texas A&M germplasm and obsolete US cultivars. The prediction error appears to be distributed evenly in both groups, as evidenced by the dispersal around the line, however; the Texas A&M data is more distributed across the x axis and this may be the cause of the higher prediction accuracy for these genotypes.

Population structure analysis was determined using the marker data and performing fastStructure on the marker data for the full population. A K of one was identified as the optimal K in using both the simple and logistic priors' argument, indicating no subclasses were identified and therefore no population structure was detected using the marker data. To look at the effects on prediction accuracy with varying degrees of population structure, an inbred substitution analysis was conducted, and the results of this analysis are shown in Table 5.1. For UHML, the prediction accuracy for the genomic prediction analysis on the obsolete US cultivars with no substitution was 0.55. This was similar to the prediction accuracies of populations where 5, 10, 15, and 20 percent of the obsolete US improved cultivars where substituted with the Texas A&M germplasm with prediction accuracies for the obsolete US cultivars of 0.56, 0.56, 0.54, and 0.54 respectively. For strength, the

prediction accuracy for the genomic prediction analysis on only the obsolete US cultivars was 0.55. The results of the inbred substitution analysis for strength showed similar results as UHML, as prediction accuracies for the 5,10, 15, and 20 percent substitutions were similar for the obsolete US cultivars at 0.57, 0.57, 0.57, and 0.55. These prediction accuracies are similar to the prediction accuracies for these cultivars in the genomic prediction model for the full population of 174 genotypes at 0.56 for UHML and 0.55 for strength. The marker data were evaluated using fastStructure on the population that had 20% of obsolete US improved cultivars substituted with the Texas A&M bidirectional inbreds to see if clear population structure was detectable in the molecular data. The optimal K was identified as one when using both simple and logistic priors, indicating no population structure detectable using the marker data. This indicates that population structure doesn't appear to effect prediction accuracy, if the structure is not detectable in the marker data. Genomic prediction only looks at marker effects to establish predicted values. Previous research has shown that prediction accuracy can be diminished due to population structure (Wientjes et al., 2013; Habier et al., 2010), as linkage disequilibrium of trait alleles with marker alleles can confound the analysis. This doesn't appear to be the case in this study; as structure was clearly present according to the phenotypic data, but was not detectable by the marker data. By squaring the prediction accuracy, the variation explained by the marker data can be calculated. The variation explained for the obsolete cultivars from the genomic prediction analysis on the whole population ranged between .06 to .31 for the various fiber traits, which is much less than the reported heritability for these fiber traits (Braden and Smith, 2004; Dabbert et al., 2017; Hugie et al., 2017; Zeng and Bechere, 2017; Ulloa, 2006).

Thus, there is likely much genetic variation that is unaccounted for by the markers used in this study, with some of this unaccounted for variation likely nested within population structure, which is also not accounted for by the markers used in this study.

The results of using the HVI and AFIS parameters to model yarn work to break are shown in Figure 5.3. The highest mean predictability for the 100 bootstrap iterations for each number of traits added by order of largest effect on OOB is 0.92 at five traits. The lowest variation for these predictabilities was 0.72 $x10^{-3}$ also at five traits. The five traits with the largest effect on the OOB in order are AFIS fineness, HVI strength, HVI elongation, HVI uniformity, and AFIS standard fineness. Using the random forest model with these five traits, predictions were made for all individuals in the whole population for yarn work to break. Genomic prediction analysis was performed, and the results are shown in Figure 5.4. Prediction accuracies is largest for the whole population for the same reasons as discussed earlier from the combination of two distinct groups that have more spread than the individual groups. Prediction accuracy for the whole population, the Texas A&M cultivars, and the obsolete US improved cultivars, are 0.66, 0.42, and 0.36 respectively. It is too expensive to conduct spin test for large populations when cost per sample is approximately 130 dollars, and considering the large amount of land needed to produce enough fiber for spinning. This cost prevented the replication of spinning data in this analysis, and is certainly prohibitive to using yarn data in normal selection cycles of a breeding program. There is induced error involved in making a prediction on a prediction; however, the prediction accuracy for yarn work to break of .924 from using HVI and AFIS parameters allowed for confidence in moving forward in the analysis with genomic

prediction. The predicted yarn values for all genotypes calculated from the HVI and AFIS index data were used in the genomic prediction analysis. The accuracy of prediction was determined not by correlating the genomic prediction to the yarn prediction, but to the empirical data from the College Station trial from which spinning data were obtained. This allowed for the error involved from the prediction of the yarn work to break to be induced into the determination of the genomic prediction accuracy. Still, this analysis showed that the marker data was able to explain some of the genetic variation in yarn work to break. Yarn work to break is too expensive for use in breeding programs for selection, and therefore alleles for this trait cannot be directly selected; however, this technique does allow for some degree of selection for these alleles.

5.4. Conclusion

The development of genomics has given breeders new tools to help move favorable alleles forward through selection. This study used new tools available to the cotton breeding community to evaluate the application of genomic prediction for fiber quality traits. HVI is a fiber bundle quality phenotyping method that has been implemented relatively recently in the scope of US cotton improvement efforts, and AFIS is an individual fiber quality phenotyping method that is not incorporated at many levels in breeding programs due to cost. Many genotypes in the USDA's obsolete germplasm collection have never been phenotyped using either of these methods. Using the obsolete cultivars from this analysis as a training population, a breeder can revisit this collection to identify individuals with potentially beneficial fiber quality alleles. This study showed that genomic prediction is

effective in selecting for yarn quality alleles, which may have application as phenotyping for yarn quality is much more expensive than genotyping.

Although this study determines that genomic prediction was successful at identifying some of the genetic variation for fiber quality, it was less than reported in the literature through phenotypic heritability studies. It was shown that although in this population there was clear population structure evident through phenotype, this structure was undetected in the marker data. The prediction accuracy was not diminished by adding varying levels of structure, which adds to the evidence that many favorable alleles are not represented in the marker platform. As marker platforms develop, this is expected to change.

Table 5.1. Genomic prediction accuracies of obsolete cultivars with varying levels of substitution with Texas A&M's bidirectional inbreds. Obsolete US cultivars were replaced at varying percentages of population size with germplasm from Texas A&M that distributed a clear population structural difference than the obsolete cultivars determined phenotypically. Prediction accuracies are given for the obsolete US cultivars. In the last column on the right prediction accuracies are given for obsolete cultivars from the genomic prediction analysis of the whole population used in this study of 174 genotypes.

Traits	Obsolete Only	5%	10%	15%	20%	Whole Population
UHML	0.545	0.564	0.561	0.549	0.535	0.558
Strength	0.547	0.570	0.568	0.570	0.550	0.550

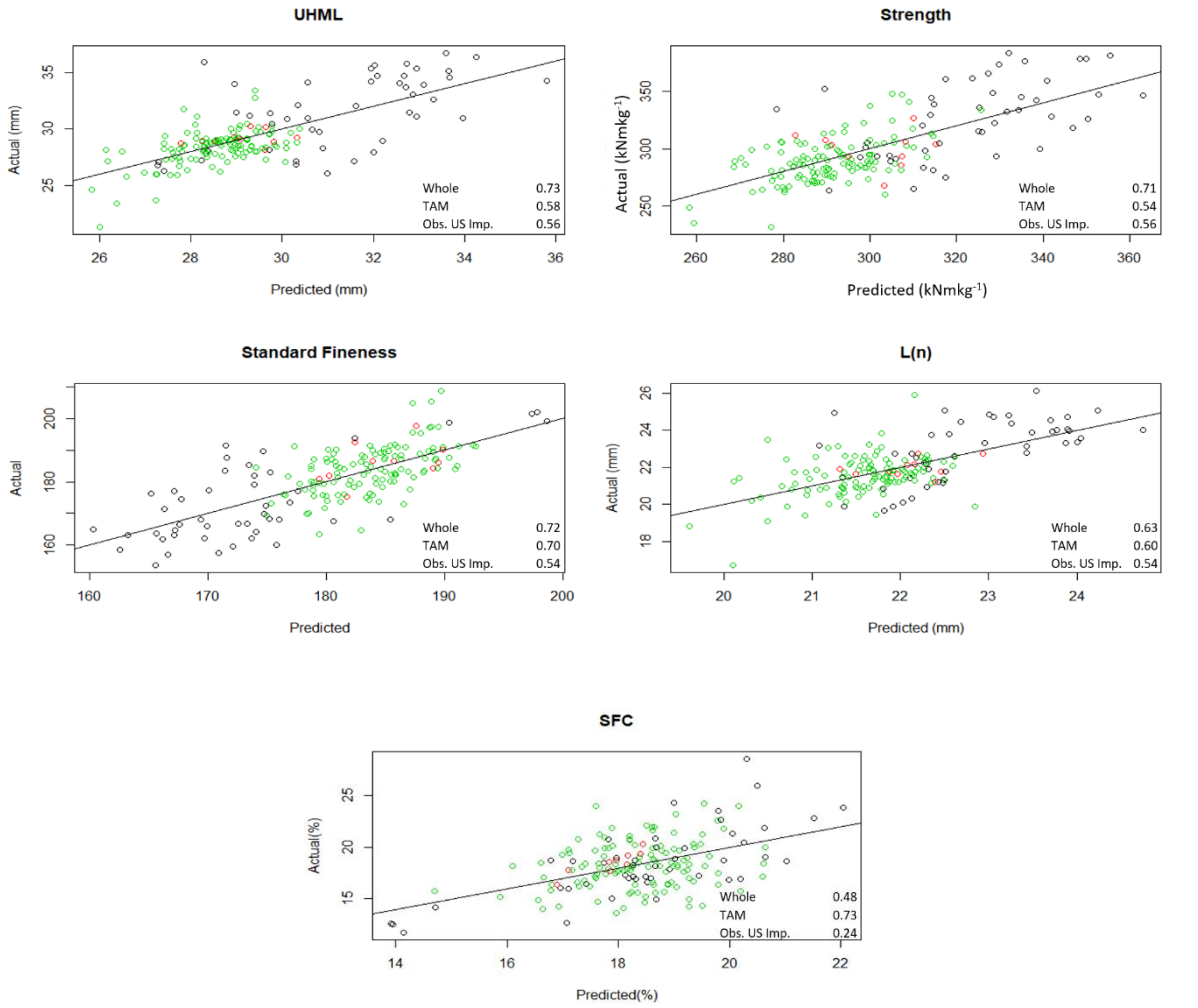


Figure 5.1. Prediction accuracies for cotton fiber traits. Line drawn is the slope for correlation of 1 between predicted value and actual value. Prediction accuracy is given as correlation between predicted value and actual value for the full population (Whole), only the genotypes from Texas A&M Cotton Improvement Lab, and only the obsolete US improved cultivars from the results of the genomic prediction analysis for the full population (obs. US Imp). The green points correspond to the obsolete US improved cultivars, the black points correspond to the genotypes from Texas A&M Cotton Improvement Lab, and the red points correspond to the 10 current commercial cultivars.

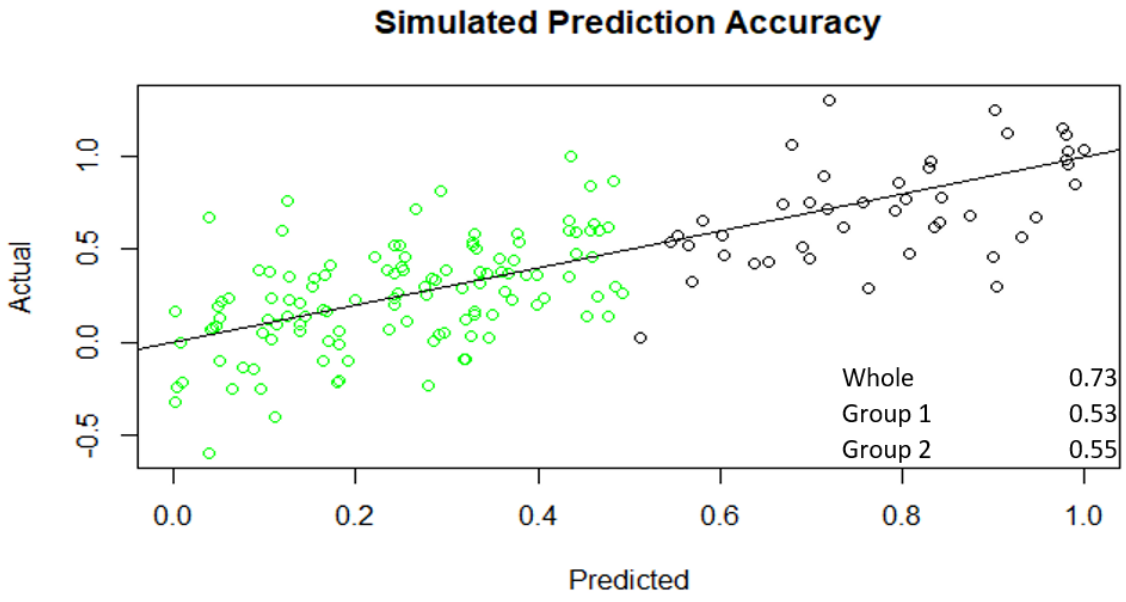


Figure 5.2. Correlations for simulated predicted and actual data showing that the correlation for the whole population is increased when there are two different phenotypic groups that spread the data out more. Group one is represented by the green points and group 2 is represented by the black points. Correlations between predicted and actual simulated data is given in the bottom right of the plot.

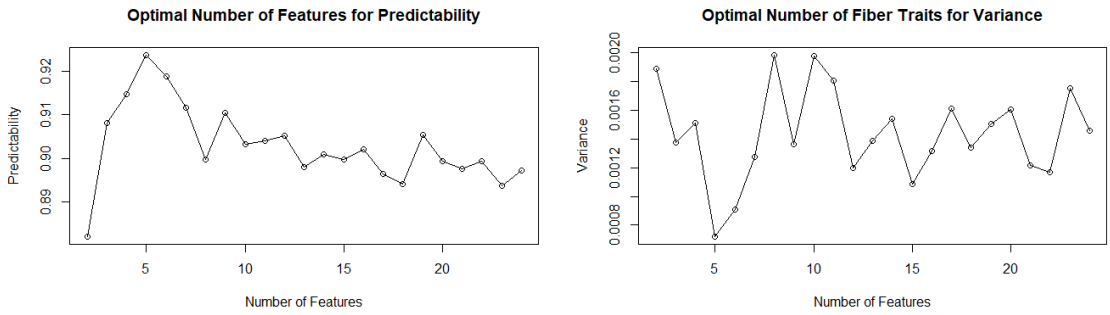


Figure 5.3. Results from random forest analysis to model yarn work to break from HVI and AFIS parameters. The left plot shows mean predictability from the 100 bootstrap iterations as determined by correlation of predicted values to actual values for the addition of each parameter ordered by largest effect on the OOB. The right plot shows the variation of the predictabilities from the 100 bootstrap iterations for the addition of each parameter ordered by largest effect on the OOB.

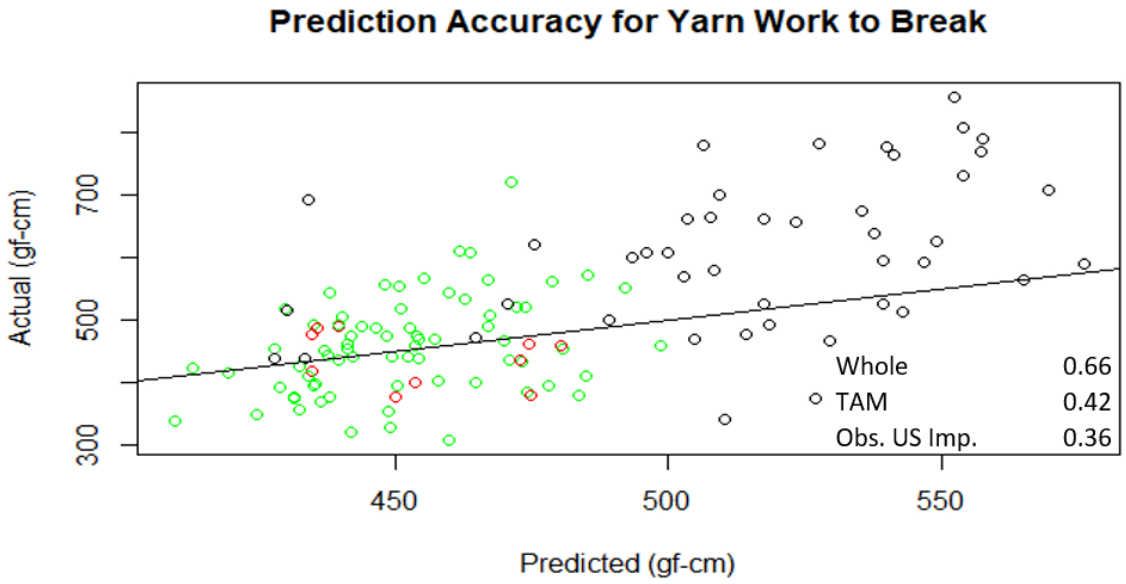


Figure 5.4. Prediction Accuracy for yarn work to break. Line drawn is the slope for correlation of 1 between predicted value and actual value. Prediction accuracy is given as correlation between predicted value and actual value for the full population (Whole), only the genotypes from Texas A&M Cotton Improvement Lab (TAM), and only the obsolete US improved cultivars from the results of the genomic prediction analysis for the full population (obs. US Imp). The green points correspond to the obsolete US improved cultivars, the black points correspond to the genotypes from Texas A&M Cotton Improvement Lab, and the red points correspond to the 10 current commercial cultivars.

5.5. References

Braden, C.A., and C.W. Smith. 2004. Fiber length development in near-long staple upland cotton. Crop Sci. 44:1553-1559.

Bragg, C.K., and F.M. Shofner. 1993. A rapid, direct measurement of short fiber content. Text. Res. J. 63:171-176.

Dabbert, T.A., D. Pauli, R. Sheetz, and M.A. Gore. 2017. Influences of the combination of high temperature and water deficit on the heritabilities and correlations of agronomic and fiber quality traits in upland cotton. Euphytica 213:UNSP 6.

Faulkner, W.B., E.F. Hequet, J. Wanjura, and R. Boman. 2012. Relationships of cotton fiber properties to ring-spun yarn quality on selected high plains cottons. Text. Res. J. 82:400-414.

Habier, D., J. Tetens, F. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genetics Selection Evolution 42:5.

Hugie, K.L., C.W. Smith, K.S. Joy, and D.C. Jones. 2017. Divergent selection for fiber length and bundle strength and correlated responses in cotton. Crop Sci. 57:99-107.

Joy, K., C.W. Smith, E. Hequet, S. Hague, P.S. Thaxton, and C. Souder. 2010. Fiber properties and mini-spun yarn performance of extra long staple upland cotton. J.Cotton Sci 14:82-90.

Liaw, A., and M. Wiener. 2002. Classification and regression by randomForest. R News 2:18-22.

Meuwissen, T., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829.

Smith, C.W., C.A. Braden, and E.F. Hequet. 2009. Generation mean analysis of near-long-staple fiber length in TAM 94L-25 upland cotton. Crop Sci. 49:1638-1646.

Thaxton, P.M., C.W. Smith, and R. Cantrell. 2005. Registration of 'Tamcot 22' high-yielding upland cotton cultivar. Crop Sci. 45:1165.

Thorwarth, P., E.A.A. Yousef, and K.J. Schmid. 2018. Genomic prediction and association mapping of curd-related traits in gene bank accessions of cauliflower. G3-Genes Genomes Genetics 8:707-718.

Ulloa, M. 2006. Heritability and correlations of agronomic and fiber traits in an okra-leaf upland cotton population. Crop Sci. 46:1508-1514.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414-4423.

Wientjes, Y.C.J., R.F. Veerkamp, and M.P.L. Calus. 2013. The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. Genetics 193:112.

Williams, G.F., and J.M. Yankey. 1996. New developments in single fiber fineness & maturity measurements. p. 1287. *In* Proceedings of the 1996 Cotton Beltwide Conference. Proceedings, Nashville, TN. January 9-12 1996. Cotton Incorporated, Memphis, TN.

Yu, X., X. Li, T. Guo, C. Zhu, Y. Wu, S.E. Mitchell, K.L. Roozeboom, D. Wang, M.L. Wang, G.A. Pederson, T.T. Tesso, P.S. Schnable, R. Bernardo, and J. Yu. 2016.

Genomic prediction contributing to a promising global strategy to turbocharge gene banks. Nature Plants 2:16150.

Zeng, L., and E. Bechere. 2017. Correlated selection responses of fiber properties measured by high volume instrument and advanced fiber information system in upland cotton. Euphytica 213:278.

Zhang, M., Y.H. Wu, M.K. Lee, Y.H. Liu, Y Rong, T.S. Santos, C. Wu, F. Xie, R.L. Nelson, H.B. Zhang. 2010. Numbers of genes in the NBS and RLK families vary by more than four-fold within a plant species and are regulated by multiple factors. Nucleic Acids Res. 19:6513-6525

6. CONCLUSIONS

This dissertation research indicated that genetic gains for both HVI and AFIS traits have been made in the past 100 years of breeding effort, even though selection for AFIS traits such as maturity, fineness, standard fineness, and immature fiber content is not widely implemented and was not available until after 1990. It is likely that gains in these traits are from correlation with traits such as length and strength in which there has been direct selection pressure, indicating there is potential standing variation for these traits that is unexploited in the obsolete US cultivar collection. It was determined that there is an increase in gains for strength that occurs around the 1940s. This supports the conclusions of Green and Culp (1990), Bowman and Gutierrez (2003), and Bowman et al. (2006) that Beasley's Triple Hybrid (Beasley 1940) contributed to the breaking of negative linkage of fiber strength with yield around this same time.

This work determined that fiber quality traits are highly stable across environments, reaffirming the findings of both Campbell et al. (2012) and Ng et al. (2013) who reported that GxE for fiber quality traits is more significant for magnitude rather than rank changes. Genomic prediction is applied to the environment in which the training population is evaluated in. If traits evaluated are unstable across environments, this will diminish predictability. Fiber quality improvements were apparent regardless of selection environment.

Finally, this work determined that genomic prediction was successful in identifying some of the genetic variation for fiber quality; however, it was less than reported in

phenotypic heritability studies by Braden and Smith (2004), Dabbert et al. (2017), Hugie et al. (2017) Zeng and Bechere (2017), and Ulloa (2006). The molecular marker platform used in this study didn't adequately account for all the genetic variation. The Texas A&M cotton material was determined to be superior in terms of fiber quality compared with the obsolete US improved cultivar collection. The addition of this germplasm into the obsolete US improved cultivar collection to generate clear population structure was identifiable phenotypically, but not molecularly. This study determined that genomic prediction is effective in selecting for yarn quality alleles, which may have application as phenotyping for yarn quality is much more expensive than genotyping. As molecular marker technologies continue to develop and better explain genetic variation, genomic prediction will become an increasingly valuable tool to breeders.

6.1. References

Beasley, J.O. 1940. The origin of american tetraploid *Gossypium* species. Am. Nat. 74:285-286.

Bowman, D. 2006. Pedigrees of upland and pima cotton cultivars released between 1970 and 2005.

Bowman, D.T., and O.A. Gutiérrez. 2003. Sources of fiber strength in the US upland cotton crop from 1980 to 2000. J.Cotton Sci 7:164-169.

Braden, C.A., and C.W. Smith. 2004. Fiber length development in near-long staple upland cotton. Crop Sci. 44:1553-1559.

Campbell, B.T., P.W. Chee, E. Lubbers, D.T. Bowman, W.R. Meredith Jr., J. Johnson, D. Fraser, W. Bridges, and D.C. Jones. 2012. Dissecting genotype x environment interactions and trait correlations present in the pee dee cotton germplasm collection following seventy years of plant breeding. Crop Sci. 52:690-699.

Dabbert, T.A., D. Pauli, R. Sheetz, and M.A. Gore. 2017. Influences of the combination of high temperature and water deficit on the heritabilities and correlations of agronomic and fiber quality traits in upland cotton. Euphytica 213:6.

Green, C.C., and T.W. Culp. 1990. Simultaneous improvement of yield, fiber quality, and yarn strength in upland cotton. Crop Sci. 30:66-69.

Hugie, K.L., C.W. Smith, K.S. Joy, and D.C. Jones. 2017. Divergent selection for fiber length and bundle strength and correlated responses in cotton. Crop Sci. 57:99-107.

Ng, E., K. Jernigan, W. Smith, E. Hequet, J. Dever, S. Hague, and A.M.H. Ibrahim. 2013. Stability analysis of upland cotton in texas. Crop Sci. 53:1347-1355.

Ulloa, M. 2006. Heritability and correlations of agronomic and fiber traits in an okra-leaf upland cotton population. Crop Sci. 46:1508-1514.

Zeng, L., and E. Bechere. 2017. Correlated selection responses of fiber properties measured by high volume instrument and advanced fiber information system in upland cotton. Euphytica 213:278.