NEW TOOLS FOR DEVELOPING IMPROVED SORGHUM HYBRIDS

A Thesis

by

DANIEL SHAW CROZIER

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

MASTER OF SCIENCE

Chair of Committee,	William L. Rooney
Committee Members,	Robert R. Klein
	Oscar Riera-Lizarazu
Head of Department,	David D. Baltensperger

August 2020

Major Subject: Plant Breeding

ABSTRACT

The genetic yield potential in grain sorghum [*Sorghum bicolor* (L.) Moench] hybrids has increased at a slower rate than other cereal crops. Advances in new technology provide opportunities for breeders to enhance selection accuracy and throughput efficiency of new germplasm to bolster rates of genetic gain. In this thesis, Genotyping-By-Sequencing was used to analyze the structure of heterotic groups in sorghum and access the relationship between the genetic similarity of parental lines and heterosis. Three distinct groups of germplasm in the Texas A&M sorghum breeding program were found through K-means clustering that closely aligned with functional classification as B-lines, R-lines, and forage lines. Forage lines exhibited the greatest range of genetic diversity followed by R-lines, then B-lines. Significant heterosis was observed for grain yield, plant height, days to flower, and panicle exsertion; yet, estimates of genetic similarity were not a good predictor of heterosis or hybrid performance amongst elite Texas A&M sorghum inbred lines. However, some parental inbred performance measurements may be predictive of hybrid performance. Additionally, in this thesis, a phenotyping pipeline was developed utilizing CT imaging to quantify three-dimensional structural characteristics from grain sorghum caryopses which can then be related to end-use quality. It was possible to accurately classify 19 sorghum genotypes based on CT-derived estimates of embryo volume, endosperm hardness, endosperm texture, endosperm volume, pericarp volume, and kernel volume.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. William L. Rooney, for providing me the opportunity to come to Texas A&M University and not only providing professional guidance, but also the many conversations we've had on long trips and around the lunch table that have challenged my world views. I would also like to thank my committee members, Dr. Robert R. Klein and Dr. Oscar Riera-Lizarazu, for their guidance and support throughout the course of this research. Additionally, I would like to thank friends, colleagues, and the faculty at Texas A&M university who aided in my research and made my time here a great experience. Finally, I would be remiss not to acknowledge the support my parents, Carl and Teresa Crozier, have given me through the years.

CONTRIBUTORS AND FUNDING SOURCES

This work was supervised by a thesis committee consisting of Dr. William L. Rooney from the Department of Soil and Crop Sciences at Texas A&M University, Dr. Robert R. Klein from the USDA-ARS Crop Germplasm Research Unit, and Dr. Oscar Riera-Lizarazu from the Department of Horticultural Sciences at Texas A&M University.

With respect to chapter 2, experimental design and phenotypic data were provided by Dr. Leo Hoffmann Jr. and Dr. William L. Rooney and support for crop management and phenotypic data collection were provided by staff and students in the Sorghum Breeding and Genetics Lab at Texas A&M University. Genotypic data was provided by Dr. Leo Hoffmann Jr. and support was provided by Dr. Patricia E. Klein.

With respect to chapter 3, CT scans were performed in coordination with Annie Hedman at the Texas A&M Cardiovascular Pathology Lab. Germplasm was provided by Dr. William L. Rooney and support for crop growth and harvest was provided by staff and students in the Sorghum Breeding and Genetics Lab at Texas A&M University.

All other work for completion of the thesis was completed by the student independently.

With respect to chapter 2, this work was made possible in part by the United Sorghum Checkoff Program under Grant Number C1001-12. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the United Sorghum Checkoff Program. Further support of graduate study was supported by

funding from the Borlaug-Monsanto Chair for Plant Breeding and International Crop Improvement.

NOMENCLATURE

BLUE	Best Linear Unbiased Estimator
CMS	Cytoplasmic-nuclear Male Sterility
CT	X-ray Computed Tomography
EBLUP	Estimated Best Linear Unbiased Predictor
GBS	Genotyping-By-Sequencing
HP	High-Parent
HPH	High-Parent Heterosis
HSD	Honest Significant Difference
LSMeans	Least Square Means
MP	Mid-Parent
MPH	Mid-Parent Heterosis
r	Pearson's Correlation Coefficients
R	Repeatability
R^2	Percent Variation Explained
REML	Restricted Maximum Likelihood Analysis
RFLP	Restriction Fragment Length Polymorphisms
RMSE	Root Mean Square Error
SKCS	Single Kernel Characterization System
SNP	Single Nucleotide Polymorphism
SSR	Simple Sequence Repeats

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

1. INTRODUCTION

Sorghum [*Sorghum bicolor* (L.) Moench] is a morphologically diverse crop grown worldwide for a variety of end-use products. In the United States and many countries, sorghum is primarily cultivated as a hybrid crop largely for animal feed markets. Regardless of end-use, sorghum improvement programs have long selected for enhanced grain yield with modest success. Several recent studies estimate the rate of genetic gain in grain yield for sorghum hybrids to be less than that of other major US field crops (Gizzi & Gambin, 2016; Pfeiffer et al., 2019). Modest gains in yield are attributed to several factors including the shifting of production to more marginal environments, reduced inputs, and a reduction in public and private sector research expenditures. Regardless of the causes, the rate of genetic improvement must increase if sorghum is to remain a major agronomic commodity.

In addition to increasing yield, manipulating and improving grain quality components are another important factor in keeping sorghum commercially viable. Recent focus shifted toward high-end specialty grain sorghums. This is evidenced by the development of niche markets that use sorghum products such as: popped sorghum, gluten free flours, alcoholic beverages, high antioxidant cereals, nutraceutical additives, packing materials, bacterial substrates, natural colorants, and biofuels (Aruna & Visarada, 2019). Consequently, varied end-use products will require tailored hybrids with specific grain structural and compositional characteristics.

To meet market demands, plant breeders continuously search for new tools and novel uses for old tools to improve throughput and selection efficiency. This study seeks

to examine two such tools and their usefulness in a sorghum breeding program. The first being applications of Genotyping-By-Sequencing (GBS) data to access the relationship between heterotic groups and predict heterosis in sorghum. The second being the development of protocol for analysis of endosperm texture in sorghum grain using X-ray computed tomography (CT).

2. PREDICTING HETEROSIS IN GRAIN SORGHUM HYBRIDS USING SEQUENCE-BASED GENETIC SIMILARITY ESTIMATES*

2.1. Introduction

Sorghum [*Sorghum bicolor* (L.) Moench] is a cereal grain crop grown worldwide for food, animal feed, and bioenergy production. In the USA and many other countries, sorghum is primarily cultivated as a hybrid crop for animal feed markets. Sorghum improvement programs have long selected for enhanced grain yield with modest success. Recent studies estimate the rate of genetic gain in grain yield for sorghum hybrids to be 0.008 t ha^{-1} yr^{-1}, which is less than that of other major US field crops (Gizzi & Gambin, 2016; Pfeiffer et al., 2019). Modest gains in yield are attributed to several factors, including the shifting of production to more marginal environments, reduced inputs, and a reduction in public and private sector research expenditures. Regardless of the causes, the rate of genetic improvement must increase if sorghum is to remain a major agronomic commodity.

Commercial hybrid sorghum seed production relies on the three-line cytoplasmic-nuclear male sterility (CMS) system as first described by Stephens and Holland (1954). In this system, the process to develop parental inbreds is both costly and time-consuming, especially developing pairs of A/B parental lines. As with any hybrid breeding program, numerous new inbred lines are tested for general combining ability

*This is an Author's Original Manuscript of an article published by Taylor & Francis in Journal of Crop Improvement on April 2, 2020, available online at the Taylor & Francis Ltd web site: https://www.tandfonline.com/doi/full/10.1080/15427528.2020.1748152

and then specific hybrid combinations with commercial potential must be identified. Methods capable of predicting hybrid sorghum performance should increase the effectiveness of elite hybrid evaluation within the same size breeding program and thereby increase the rate of genetic gain to levels observed in other crops (Basnet et al., 2019; Voss-Fels et al., 2019).

Heterosis, or hybrid vigor as first described by Shull (1908, 1914), is the phenomenon observed when F_1 hybrids exhibit improved performance over either inbred parent. In sorghum, heterosis is an important component of hybrid performance that manifests in hybrids that are earlier maturing, slightly taller, and higher yielding with greater yield stability than inbred lines (Axtell et al., 1999; Duvick, 1999; Mindaye et al., 2016; Quinby, 1963). Mid-parent heterosis (MPH) for grain yield in sorghum is substantially lower than in maize (Duvick, 2005; Pfeiffer et al., 2019). The lower heterosis in sorghum is thought to be caused by the natural self-pollination in the crop and the absence of heterotic groups from the initial deployment of hybrids (Duvick, 1999; Pfeiffer et al., 2019). The latter factor is supported by the increase in heterosis in grain sorghum breeding programs over time (Menz et al., 2004; Pfeiffer et al., 2019). This demonstrates the ability to select and improve heterosis in sorghum for future genetic gains.

While the underlying genetic mechanisms that govern heterosis are not fully understood (Schnable & Springer, 2013), a higher level of heterosis has been reported between parents that are more genetically distant than closely related parents (Amelework et al., 2017; Hallauer, 1999; Moll et al., 1965; Stupar et al., 2008; Wegary

et al., 2013; T. Zhang et al., 2010). However, other studies are less definitive and report little to no correlation between genetic similarity and heterosis depending on the molecular marker type, trait, and crop (Flint-Garcia et al., 2009; Nie et al., 2019; Wu et al., 2013). To date, relatively few studies have assessed the relationship between genetic similarity and heterosis in sorghum. Jordan et al. (2003) reported significant correlations between genetic similarity and hybrid yield, height, and maturity. Mindaye et al. (2016) reported a significant correlation between MPH for grain yield and genetic similarity, but not in all environments. In contrast, Amelework et al. (2017) found no significant relationship between genetic similarity and grain yield heterosis in sorghum.

In highly genetically divergent parental material, heterotic expression is decreased presumably due to the beginnings of intraspecific genetic incompatibility (Moll et al., 1965; Wei & Zhang, 2018). Alternatively, genetically similar material will not manifest as much hybrid vigor. Consequently, there is a theorized optimal mating distance between parents to maximize heterosis and hybrid performance (Wei & Zhang, 2018). Furthermore, in some parental material exhibiting extreme genetic divergence, one parent brings a yield drag or other unacceptable agronomic traits, such as height or maturity, that are manifested in the hybrid. For these reasons, panels of elite parental material that produce agronomically acceptable hybrids as opposed to exotic or unadapted germplasm with greater magnitude of genetic divergence are more relevant to breeding programs.

Sorghum breeding programs have traditionally assessed genetic relatedness through analysis of ancestral relationships and morpho-anatomical traits. The

development of molecular markers, such as restriction fragment length polymorphisms (RFLPs) and simple sequence repeats (SSRs), has allowed calculation of genetic similarity with greater accuracy (Powell et al. 1996; Menz et al. 2004). However, when using RFLP and SSR markers, estimates of genetic similarity are greatly affected by the number and distribution of markers used (Powell et al., 1996). In contrast, the cost-effectiveness and dense uniform distribution of single nucleotide polymorphism (SNP) markers make them an ideal candidate to measure genetic relatedness with high throughput efficiency. To this end, Morishige et al. (2013) developed a Genotyping-By-Sequencing (GBS) pipeline for sorghum that employs methylation-sensitive enzymes that cut in gene-rich regions of the genome. This system has been effectively used to determine the genetic relatedness of various sorghum genotypes, both in diverse germplasm pools (Horne, 2019; Klein et al., 2015) and in highly related germplasm (Patil et al., 2019).

In the present study, GBS was employed to assess the genetic relationship among germplasm within the Texas A&M sorghum breeding program, and subsequently, whether genetic similarity estimates were predictive of heterosis or hybrid performance among elite inbreds. The ultimate goal of this work is to determine whether genetic similarity estimates alone, or when used in conjunction with additional information (inbred line performance), can be useful for prediction models in hybrid grain sorghum development.

2.2. Materials and methods

2.2.1. Plant material

A representative collection of germplasm in the Texas A&M sorghum breeding program, consisting of 435 sorghum inbred lines, was evaluated in this study. From this, a subset of 24 elite grain-type inbred lines, consisting of publicly released and experimental germplasm from the Texas A&M sorghum breeding program, was further evaluated. This subset of elite lines was chosen to reflect the diversity of elite material with commercial potential developed within the Texas A&M sorghum breeding program. Selected parental lines were divided into two groups, 15 female parents (A/B-lines) and nine male parents (R-lines), and 52 F_1 hybrids were made by crossing these parents.

2.2.2. Experimental details

In the spring of 2015, hybrids and parental lines were grown in four Texas locations: College Station, Lubbock, Plainview, and Vega. The experimental design in each location was a randomized complete block design with three replications. An experimental unit was a two-row plot that was between six and ten meters in length. Row spacing ranged from 0.67 to 1.0 m depending on location. Standard agronomic practices for optimum sorghum production were followed, including judicial use of pesticides, fertilizer, and irrigation as needed.

2.2.3. Data collection

In each environment, phenotypic data were collected on days to flowering, plant height, panicle exsertion, and grain yield. Days to flower was the number of days from planting

to the date on which 50% of plants in a plot had reached half-bloom. Panicle exsertion was measured at maturity as the distance from the collar of the flag leaf to the base of the panicle. Plant height was recorded at maturity as the distance from the soil surface to the tip of the panicle. Grain yield was measured using a plot combine, and grain weights are presented in $kg \bullet ha^{-1}$ with adjustments to 14% moisture content.

2.2.4. Genotypic analysis

To assess genetic diversity among the parental lines, GBS (Morishige et al., 2013) was conducted on the panel of 435 inbreds. Twenty seeds from each line were germinated in Sunshine MVP growing media (Sun Gro Horticulture, Agawam, MA) in a greenhouse for 14 days under natural sunlight supplemented with sodium halide lights. Total genomic sequence-quality DNA was extracted from leaf tissue (~12 seedlings) from each parental line with the FastPrep FP120 instrument (MP Biomedicals, Santa Ana, CA) used in conjunction with the FastDNA Spin Kit (MP Biomedicals, Santa Ana, CA) according to the manufacturer's protocol. Purified DNA was quantified fluorometrically using a Qubit Fluorometer (Invitrogen, Carlsbad, CA). Illumina template libraries were prepared by digesting each sample with *Ngo*MIV restriction enzyme as previously detailed by Burrell et al. (2015). Single-end sequencing of the template was performed on an Illumina HiSeq 2500 using standard Illumina protocols at the Texas A&M AgriLife Genomic and Bioinformatics Services. The sequences obtained were processed through a series of custom Perl and Python scripts to discard sequences lacking the partial *Ngo*MIV restriction site and 12-bp barcode, sort sequences corresponding to each genotype into separate files, and trim the 12-bp barcode from the genomic sequences.

The processed reads were aligned to the sorghum reference genome of genotype B.Tx623 (Sbicolor v1.4, Phytozome) (Paterson et al., 2009) within the CLC Genomics Workbench (Qiagen, Hilden, Germany), and reads that mapped to more than one region of the reference genome were removed. For variant detection in the CLC Genomics Workbench, the parameter settings included a maximum gap and mismatch count of three, a minimum quality of the SNP base of 20, a minimum average quality of the nucleotides surrounding the SNP of 15, a neighborhood radius of five and a minimum read coverage of nine. The mapping and variant files were exported as SAM and comma-separated-value (csv) formats, respectively, and further processed using custom scripts written in Perl and Python. Genomic positions where base calls were scored in at least 25% of the parental lines were retained, and missing data were imputed using fastPHASE (Scheet & Stephens, 2006). Following imputation, 69,874 SNPs, with a minor allele frequency of greater than 5%, were retained for further use. Markers spanned all 10 chromosomes, with greater density near the ends of chromosomes where gene density is highest. Fewer markers were present in the repeat-rich heterochromatic regions surrounding the centromeres. Genetic similarity among the parental lines was calculated using Nei's genetic distance matrix in PowerMarker V3.25 (Liu & Muse, 2005).

2.2.5. Genetic structure analysis

K-means clustering was done on 435 inbred lines in JMP (Version *15.0.0*. SAS Institute Inc., Cary, NC) based on genetic similarity between lines. Iterative clustering was done from 2 to 10 clusters and peak cubic clustering criterion was observed with three

clusters. Squared Euclidean distances for genotypes to their respective cluster center, based on functional classification, were calculated. The three derived clusters were visualized using principal component analysis based on genetic similarity between lines. The same clustering analysis was performed with the subset of 24 elite grain-type lines chosen for hybrid evaluation. Two clusters, as determined by the peak cubic clustering criteria, were similarly visualized using principal component analysis. Tukey-Kramer's HSD test was used to compare differences between groups using genetic similarity and squared Euclidean distances.

2.2.6. Statistical analysis

Statistical analysis was performed using JMP, with inbreds and hybrids analyzed separately. In both, outliers were removed and homogeneity of variance was checked using the Brown-Forsythe test. A combined restricted maximum likelihood analysis (REML) was fitted for all environments using the following model:

$$Y_{ijk} = u + Gen_j + Env_k + Rep_i\,(Env_k) + Gen_j \times Env_k + E_{ijk}$$

where Y_{ijk} is the trait value, u is the mean, Gen_j is the effect of the jth genotype, Rep_i is the effect of the ith replicate, Env_k is the effect of the kth environment, Rep_i (Env_k) is the effect of replicates nested within environments, $Gen_j \times Env_k$ is the genotype-by-environment interaction and E_{ijk} is the random error term. Residuals from the models were normal for all traits, except for days to flower, based on QQ plots. Transformations for non-normal data were tried using Box-Cox functions; however, no significant improvement occurred, and thus untransformed data were used in statistical analyses. Variance components were estimated considering all factors as random effects, and the

Wald test was used to determine the significance of factors. Least square means (LSMeans) were estimated considering genotype as a fixed effect and all other factors as random effects. Pearson's correlations, simple and multiple linear regression, mid-parent (MP) values, MPH, and high-parent heterosis (HPH) were calculated using LSMeans. Percent MPH and HPH were calculated using the following formulas:

$$\%MPH = [(F_1 - MP) / MP] \times 100$$

$$\%HPH = [(F_1 - HP) / HP] \times 100$$

where F_1 is the performance of the F_1 hybrid, MP is the average performance of the parents in the cross, and HP is the performance of the high parent in the cross. A one-sample *t*-test was used to assess whether heterosis was significantly different than zero. Simple and multiple linear regression models were constructed, and Mallows C_p and root mean square error (RMSE) were used for variable/model selection criteria.

2.3. Results

2.3.1. Genetic relatedness

The genetic relationship of 435 parental lines in the Texas A&M sorghum breeding program was assessed through K-means cluster analysis based on genetic similarity and visualized using the principal component analysis (Figure 1). Clustering revealed three major groups; one primarily composed of R-lines, one mostly of B-lines, and the other comprised largely of forage lines. However, little separation was present between groups and some crossover between functional classification and genetic similarity-based clustering was present. Squared Euclidean distances for genotypes to their respective cluster center, based on functional classification, were calculated and averaged for the

three clusters to assess the amount of diversity present in each. Forage lines were more diverse than both B and R clusters based on Tukey-Kramer's HSD test ($p < 0.001$) and had a mean squared Euclidean distance to the cluster center of 356.9. B-lines and R-lines did not differ significantly in diversity from one another ($p = 0.150$) and had a mean squared Euclidean distance to cluster centers of 209.7 and 241.5, respectively. An alternative analysis compared the genetic similarities for all B-lines, R-lines, and combinations between B-lines and R-Lines. The average genetic similarity among B-lines, R-lines, and between B-lines and R-lines was 0.753, 0.715, and 0.689, respectively. Differences between groups were significant, as indicated by Tukey-Kramer's HSD test ($p < 0.001$), with the most similarity being observed among the B-lines, and the least being observed between B-lines and R-lines.

Having examined the genetic relationship among a large panel of sorghum genotypes within the Texas A&M sorghum breeding program, focus shifted to a narrower set of 24 elite grain sorghum B-lines and R-lines (Figure 2). This smaller collection is representative of elite grain-type inbreds developed within the Texas A&M sorghum breeding program with commercial potential. Analysis revealed elite parental lines clustered into two groups (Figure 2). Each cluster contained exclusively B-lines or R-lines; no crossovers between functional classification and K-means clustering, based on genetic similarity, were observed. Comparing the clusters of elite B-lines and R-lines, mean squared Euclidean distances to the respective cluster centers for R lines was 19.1 and that for B-lines was 11.8. Therefore, as confirmed by Tukey-Kramer's HSD test, more genetic diversity was observed in elite R-lines than B-lines ($p = 0.003$). The

average genetic similarity among elite B-lines, R-lines, and between B-lines and R-lines was 0.782, 0.717, and 0.695, respectively. Differences between groups were significant, as indicated by Tukey-Kramer's HSD test ($p < 0.001$), with the most similarity being observed among B-lines, and the least being observed between B-lines and R-lines.

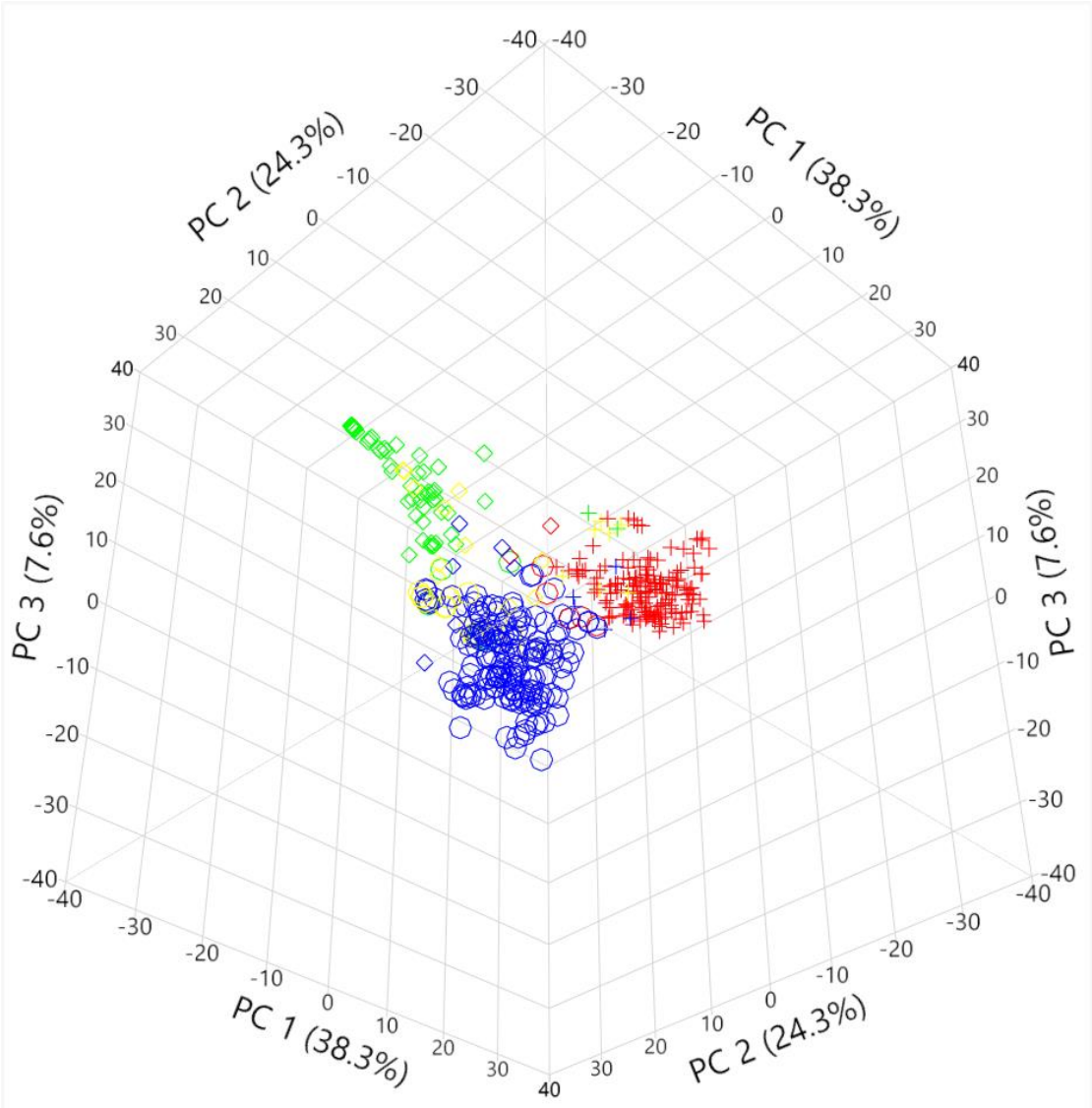

Figure 1. Principal component analysis based on genetic similarity showing the relationship between 435 sorghum inbred lines. Lines are clustered into three groups through K-means clustering which are designated by the marker shape. Lines are colored based on functional grouping: red = grain-type B-line, green = forage line, blue = grain-type R-line, and yellow = unknown classification.

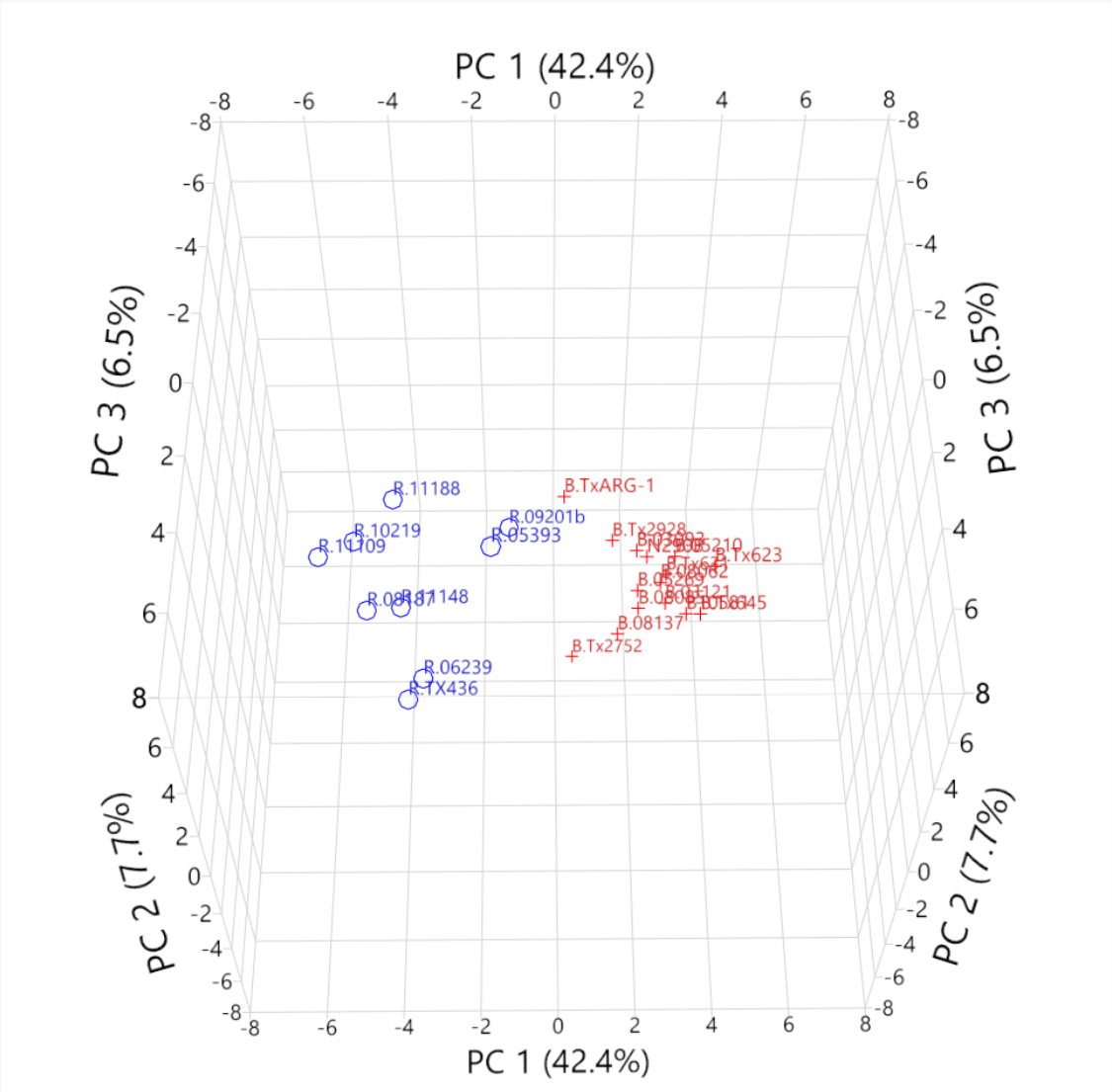

Figure 2. Principal component analysis based on genetic similarity showing the relationship between twenty-four elite grain-type sorghum inbred lines. Lines are clustered into two groups through K-means clustering which are designated by the marker shape. Lines are colored based on functional grouping: red = B-line, blue = R-line. The genotype of each line is displayed alongside its respective marker.

2.3.2. Inbred and hybrid performance

In the combined analysis across four environments, fitted separately for inbreds and hybrids, genetic variation was significant for all traits examined in both inbred and

hybrid performance trials (Table 1). In addition, significant genotype-by-environment interactions were observed for all traits in both inbreds and hybrids. Across environments, hybrids exhibited a comparative grain yield advantage of 1380.6 kg ha^{-1} (36.9%) over inbred parental lines. In addition, hybrids reached mid-anthesis 6.6% earlier, were 23.4% taller, and had 70.2% greater panicle exsertion than inbred parental lines (Table 2).

Table 1. Percent variance associated with sources of variation in the combined model fitted separately for 24 inbred parents and 52 hybrids across four environments in Texas. The Wald test was used to determine the significance of variance components.

	Inbred parents				Hybrids			
Sources of variation	HT	EX	DY	Y	HT	EX	DY	Y
Genotype	48.4***	35.5***	4.4**	3.0*	33.1**	22.6***	1.3**	4.1*
Environment	18.5	23.7	89.8	64.4	38.5	26.8	94.4	31.1
Replication	0.9	1.7	0.1	0.0	1.8	7.4	0.4	1.3
G x E	7.8**	10.6***	3.2**	9.9***	6.3***	6.5**	1.8***	23.9***
Residual	24.4	28.5	2.5	22.7	20.3	36.7	2.1	39.7
Total	100	100	100	100	100	100	100	100

Notes: HT = plant height, EX = panicle exsertion, DY = days to flower, Y = grain yield; G x E = genotype by environment interaction; Significance levels indicated by *($p < 0.05$), **($p < 0.01$), and ***($p < 0.001$).

Table 2. Summary of average performance of 52 grain sorghum hybrids and 24 inbred parental lines across four environments in Texas.

	Inbred parents			Hybrids		
Trait	Mean	SD	Range	Mean	SD	Range
Grain yield (kg•ha^{-1})	3,744	522	2,752– 4,620	5,124	656	3,377 – 6,294
Plant height (cm)	119	14	94 – 143	147	10	129 – 169
Days to flowering	67	3	59 – 73	62	2	57 – 65
Panicle exsertion (cm)	10	5	2 – 24	18	4	10 – 30

Note: SD = standard deviation.

2.3.3. Heterosis and genetic similarity

MPH and HPH were significant for all traits, as indicated by a one-sample t-test (Table 3). All traits had positive MPH and HPH, except days to flower. Negative heterosis for days to flowering indicated that hybrids matured earlier than their respective inbred parental lines.

Genetic similarity between parents used in hybrid combination ranged from 0.632 to 0.792, with a mean of 0.697 (Table 4). Significant correlations existed among the genetic similarity between parents with both MPH and HPH, but there were no other significant correlations among the genetic similarity between parents with MPH or HPH for other traits (Table 3). MPH exhibited a greater magnitude of correlation to genetic similarity between parents than did HPH. In addition, as MPH considers the genetic contribution from both parents, as opposed to only the high parent, further analysis in this study focused on MPH.

Although not significant, trends across traits indicated greater MPH was associated with crosses of more dissimilar parents. This is exemplified in Table 3, where a negative correlation can be seen between MPH for all traits (except days to flower) and genetic similarity between parents. Given the direction of MPH for days to flower, a positive correlation between MPH for days to flower and genetic similarity between parents is consistent with the aforementioned trend.

Table 3. Average percent mid-parent and high-parent heterosis across all hybrids and Pearson correlation coefficients between genetic similarities of inbred parents and observed heterosis. Nonzero heterosis was accessed using a one-sample *t*-test.

	MPH				HPH			
Trait	Mean	SD	Range	GS - MPH correlation	Mean	SD	Range	GS - HPH correlation
Grain yield	33.6***	15.2	2.8 – 67.8	-0.174	25.3***	14.8	-9.8 – 54.7	-0.102
Plant height	22.5***	5.0	13.8 – 38.3	-0.340*	14.9***	6.4	1.1 – 31.4	-0.325*
Days to flowering	-5.6***	2.5	-11.8 – 0.9	0.213	-8.2***	3.5	-15.4 – -0.3	0.154
Panicle exsertion	74.0***	47.7	14.6 – 288.9	-0.080	30.1***	32.9	-22.7 – 141.4	0.040

Notes: SD = standard deviation; GS = genetic similarity; MPH = mid-parent heterosis, HPH = high-parent heterosis; Significance levels indicated by *($p < 0.05$), **($p < 0.01$), and ***($p < 0.001$).

Table 4. Genetic similarity matrix among parental-inbred lines of the 52 examined hybrids based on Genotyping-By-Sequencing SNP data.

Parental lines	R.Tx436	R.05393	R.06239	R.08187	R.09201b	R.10219	R.11109	R.11148	R.11188
N290B	0.707		0.717						
B.Tx2752	0.697		0.687	0.643		0.632		0.667	
B.Tx2928	0.696		0.685		0.732	0.686		0.687	0.675
B.Tx623	0.716	0.792	0.705	0.688		0.691		0.719	
B.Tx631	0.678	0.699							
B.Tx645	0.719		0.702	0.691		0.673			
B.TxARG-1	0.694		0.714			0.704			
B.03093	0.715			0.686	0.720			0.702	
B.03121			0.719	0.672					
B.05210	0.699		0.703	0.689			0.676		
B.05269	0.705		0.694						
B.08062	0.691		0.706	0.713					
B.08085	0.705								
B.08137		0.704	0.692	0.680		0.670		0.686	
B.10581	0.716		0.715	0.688					

Notes: B-lines in which the sterile A-line were used as female parents are denoted in red; R-lines used as fertile male restorer lines are denoted in blue, Empty spaces indicate hybrid combinations not tested.

2.3.4. Correlations with hybrid performance

Correlations between hybrid performance and genetic similarity of parental lines were non-significant for all traits, except for panicle exsertion (Table 5). MPH for grain yield was significantly correlated with hybrid yield (r = 0.660), as was MPH for days to flower and hybrid days to flower (r = 0.482). A significant negative correlation existed between MPH for parental panicle exsertion and hybrid panicle exsertion, and no significant correlation was found between MPH for parental height and hybrid height

(Table 5). The mid-parent value for a given trait was significantly and positively correlated with hybrid performance for all respective traits. Of most interest was mid-parent grain yield, which was correlated with hybrid grain yield (r = 0.553) (Table 5).

Table 5. Pearson correlation coefficients between hybrid and average mid-parent values, mid-parent heterosis, and genetic similarity from 52 elite grain-type sorghum hybrids.

Hybrid traits	Mid-parent	MPH	Genetic similarity
Grain yield	0.553***	0.660***	-0.075
Plant height	0.863***	-0.081	-0.206
Days to flowering	0.598***	0.482***	-0.076
Panicle exsertion	0.895***	-0.379**	0.299*

Notes: MPH = mid-parent heterosis; Significance levels indicated by * ($p < 0.05$), ** ($p < 0.01$), and *** ($p < 0.001$).

2.3.5. Prediction of hybrid yield

Grain yield is the most critical component of hybrid grain sorghum performance, and the ability to predict hybrid performance is central to increasing the rate of genetic gain. To this end, genotypic and phenotypic information presented herein was used to construct simple and multiple linear regression models in an attempt to predict hybrid grain yield. Based on stepwise regression, in which minimum RMSE and Mallows C_p were used as selection criteria, the best model to predict hybrid grain yield used mid-parent grain yield, mid-parent plant height, mid-parent days to flower, and genetic similarity (Table 6). The model provided significant predictive capability, as determined by an F-test ($p <$ 0.001), with an R^2 of 0.417. In the effects test, mid-parent yield and mid-parent height

were significant. However, genetic similarity and mid-parent maturity were not significant. Thus, despite inclusion of genetic similarity in the best predictive model, this factor was not a major determinant of hybrid grain yield in this study.

Table 6. Summary of simple linear regression models and best multiple linear models using two, three, four, and five predictors as determined by stepwise regression for predicting hybrid grain yield from parental traits.

Independent variables	R^2	RMSE	C_p
MY	0.306***	553.03	7.04
MMA	0.089*	633.67	24.26
MHT	0.023	656.09	29.47
MEX	0.008	660.10	30.63
GS	0.006	661.95	30.86
MY, MHT	0.364***	534.60	4.41
GS, MY, MHT	0.385***	531.24	4.76
GS, MMA, MY, MHT	0.417***	522.68	4.22
GS, MEX, MMA, MY, MHT	0.420***	527.09	6.00

Notes: MY = mid-parent yield, MMA = mid-parent maturity, MHT = mid-parent height, MEX = mid-parent exsertion, and GS = genetic similarity, RMSE = root mean squared error, C_p = Mallows C_p; Significance levels indicated by * ($p < 0.05$), ** ($p < 0.01$), and *** ($p < 0.001$).

2.4. Discussion

The large panel of sorghum lines formed distinct groups based on their functional classification as B-lines, R-lines, or forage lines (Figure 1). Previous studies using genotypic analysis to distinguish between heterotic groups in sorghum were inconclusive. Some reported that B-lines and R-lines cluster separately (Ahnert et al.,

1996), whereas others reported a higher degree of correspondence to ancestral racial origins (Brown et al., 2011; Menz et al., 2004; Mindaye et al., 2015; Ramu et al., 2013). The necessity of using a CMS system to produce seed of commercial hybrids dictates that heterotic groups be based on the fertility reaction score of sorghum genotypes. Sorghum genotypes were detected that did not cluster based on their fertility reaction score (e.g., R-lines clustered with B-lines). The crossovers and lack of marked distance between groups are reflective of the inbreeding in foundational sorghum accessions before heterotic pools were established. When examining the subset of 24 elite parental grain inbreds, clusters of genotypes based on B-lines or R-line classification were mutually exclusive and thus, represented genetically distinct heterotic pools (Figure 2). The increase in heterosis in sorghum since the development of hybrids (Gizzi & Gambin, 2016; Pfeiffer et al., 2019), and results presented herein support that sorghum breeding efforts have resulted in the development of genetically distinct heterotic pools. However, there is less divergence among sorghum heterotic groups compared with outcrossing species, such as maize (Duvick, 2005; Niebur et al., 2004). This implies that there are opportunities for future selection and breeding of more divergent complementary groups.

In both the large sorghum panel of 435 lines and the subset of elite grain-type lines, R-lines exhibited more genetic diversity than B-lines. Menz et al. (2004) found similar results in a collection of 50 elite and historically important sorghum inbred lines. These results reflect the restrictive development process for new pairs of A/B lines and the resultant slower speed at which developed germplasm can be incorporated back into

the breeding program. Additionally, average similarity among B-lines increased (0.753 to 0.782) from the large panel to the subset of elite grain-type lines, whereas no change in average similarity was detected among R-lines. Thus, showing that the effects of the restrictive development process of new A/B-line pairs is further amplified within elite material. Maintaining robust genetic variation within both heterotic groups is advantageous for continual advances in genetic gain; therefore, targeted efforts to increase diversity in sorghum should focus on developing more elite B-lines.

Genetic variation was observed within both B-lines and R-lines in the subset of elite germplasm that facilitated selection of parents for use in hybrid combinations with varying degrees of genetic divergence. Previous studies in sorghum have shown an average genetic similarity between B-lines and R-lines of 0.73, with a range of 0.61 to 0.99 using genome-wide SNP markers (Mindaye et al., 2015), or an average genetic similarity of 0.584, ranging from 0 to 0.886 using a series of AFLP and SSR markers (Menz et al., 2004). The estimates of genetic similarity in the present study were similar, but the range of genetic similarity was lower (Table 4). This observation could be attributable to the marker coverage and type of marker systems utilized in previous studies (e.g., dominant markers, such as AFLPs and limited number of SSRs). Alternatively, the present study placed an emphasis on elite inbreds as opposed to the large numbers of unadapted landraces or historical sorghum accessions used in the aforementioned studies (Menz et al., 2004; Mindaye et al., 2015).

Significant correlations between hybrid performance and MPH for several traits confirm that heterosis is an important component in determining hybrid sorghum

performance (Table 5). Heterosis was associated with sorghum hybrids that produced more grain yield, matured faster, were taller, and had greater panicle exsertion (Table 3). This is consistent with the previously reported findings on effects of heterosis in sorghum (Axtell et al., 1999; Duvick, 1999; Mindaye et al., 2016; Quinby, 1963).

Jordan et al. (2003) and Mindaye et al. (2016) reported significant correlations between genetic similarity and grain yield, height, maturity, and MPH for grain yield in sorghum. However, Amelework et al. (2017) reported no significant relationship between genetic similarity and grain yield heterosis in sorghum. While general trends in the results of this experiment indicate the magnitude of heterosis was less between more genetically similar parents, significant correlations were not seen for most traits (Table 3). This suggests that the relationship that exists is variable and dependent on the specific genotypes within a population, and thus inferences related to the relationship between diversity and heterosis are likely fixed to those specific populations (Jordan et al., 2003; Mindaye et al., 2016). Accordingly, genetic similarity is not clearly linked to heterosis or hybrid performance between heterotic groups in the elite material tested herein.

Rooney (2004) previously reported that although significant correlations existed between the performance of the parental line *per se* and hybrid performance, the correlation ($r = 0.19$) was not strong enough for direct application in a program. However, this estimate was derived from a narrow range of genetic diversity present in recombinant inbred line populations and was based on one-parent value, as opposed to mid-parent values. In this study, a much stronger correlation ($r = 0.553$) existed between

mid-parent grain yield in parental lines and hybrid grain yield. This, combined with other phenotypic and genetic traits, could be useful in prediction models to estimate grain yield in an untested potential hybrid. Thus, the old adage, “a good inbred often makes a good hybrid” professed by some sorghum breeders of generations past may have some basis in fact.

There exists a consistent relationship between height and yield in sorghum in which taller hybrids yield more when lodging is not an issue (Rooney 2004). Supporting this, recent Texas A&M sorghum variety performance trials show a consistent positive relationship between plant height and yield and no relationship between plant height and lodging (Schnell, 2018). This means breeders can push the height envelope in sorghum to achieve higher yields. In addition, when environmental stresses, such as drought, are encountered, earlier-maturing hybrids generally produce higher yields. In the USA, because sorghum is predominantly grown in marginal dryland environments, stress is often encountered and earlier-maturing hybrids perform better under these conditions (Rooney 2004).

Using this background information, a prediction model was constructed to estimate hybrid grain yield using the phenotypic traits of the parental lines and genetic similarity between lines. In the model development process, RMSE and Mallows C_p were used to prevent over parameterization and biased estimation. Inclusion of genetic similarity data into the model increased the overall predictive capacity and fit of the model by a small amount, but it was not significant in the effect test (Table 6). This reaffirmed that genetic similarity within genotypes examined herein was not a significant

predictor of hybrid performance. Ultimately, the best model was significant (R^2 = 0.417) and there was no evidence of overfitting or biased estimation. There is potential usefulness of this model in selection of hybrid combinations and predicting testcross performance of new inbred lines, pending further validation.

2.5. Conclusion

Distinct heterotic groups are present in the germplasm contained in Texas A&M sorghum breeding program. Collective evidence presented herein and in previous studies shows that there is less diversity among sorghum B-lines than R-lines and this is likely attributable to the restrictive development process for new pairs of A/B-lines. Focusing only on a subset of elite parental lines limited the range of genetic diversity observed, but it provided an opportunity to examine the relationship between genetic similarity and heterosis within parental lines that produce elite grain hybrids. Parental inbred performance was correlated with hybrid performance and may be predictive enough to be of use in estimating grain yield in untested sorghum hybrids. However, generally there was no relationship between more genetically distant parents and heterosis or hybrid performance between the genotypes comprising the two heterotic pools. As other studies have reported genetic similarity as a component in predicting heterosis or hybrid performance in sorghum germplasm, the results collectively indicate that the inferences on these relationships are best restricted to the germplasm pools from which the estimates were obtained. To accurately predict heterosis, further research into genomic prediction models or more informative marker systems is needed. Recent advances in genomics research indicate that the information content of SNPs is highly dependent on

their genome location (e.g., open vs. closed chromatin and physical distance from gene models). Thus, future studies on genetic diversity using informative markers or haplotypes may improve the power of genetic diversity in predicting hybrid performance.

3. APPLICATIONS OF X-RAY COMPUTED TOMOGRAPHY TO ANALYZE THE STRUCTURE OF SORGHUM GRAIN

3.1. Introduction

In sorghum, breeders have primarily focused on yield improvement and stability; grain quality has been a trait of secondary importance. However, the demand for cereal grains that increase animal feed efficiency, address global malnutrition, improve food quality for human consumption, and development of niche markets for sorghum dictate the necessity for improvements in grain quality (Aruna & Visarada, 2019; Bouis & Welch, 2010; Thornton, 2010). While programs can select for improved grain quality, they must maintain grain yield and yield stability as any reduction would be detrimental to adoption and further reduce rates of genetic gain for yield. In wheat, it is possible to improve grain quality parameters without sacrificing agronomic performance, so it holds the same may prove to be true in sorghum (Anderson et al., 1998).

A sorghum caryopsis is composed of three biological components: pericarp, endosperm, and embryo (germ) (Rooney & Miller, 1981). The relative size of each component varies among genotypes and production environment but pericarp, endosperm, and embryo account for around 7%, 84%, and 9% of kernel volume, respectively (Bean et al., 2016; Bidwell et al., 1922). The pericarp is the outermost layers of a kernel and includes the pericarp, mesocarp, and testa layers. The thickness of the pericarp is associated with multiple traits which affect sensitivity to grain weathering, processing qualities (ie, decortication), and storage stability (Earp & Rooney, 1982; Glueck & Rooney, 1980). The endosperm is composed of protein and

starch and is subdivided into the aleurone layer, peripheral, hard (vitreous), and soft (floury) portions (Bean et al., 2016). Lastly, the embryo is composed of the embryonic axis and scutellum and contains protein and the majority of lipids, vitamins, and minerals found in the caryopses (Waniska & Rooney, 2001).

Studies on kernel structure in sorghum have traditionally involved hand dissection or bisectioning of kernels followed by observing the single longitudinal cross section (Bidwell et al., 1922; Glueck & Rooney, 1980; Hubbard et al., 1950; Menkir et al., 1996; Patil et al., 2019). These methods destroy the seed, offer only one cross section (typically in the middle of the kernel), and are time consuming, which limits the number of kernels that can be evaluated per sample. Because kernel components vary across cross sections (Rooney & Miller, 1981), cutting kernels in half is not representative of the whole seed resulting in biased estimates of kernel structure.

In sorghum, grain quality often depends on matching the end-use with kernel characteristics. Physical properties that affect grain quality include seed size, 100 seed weight, endosperm texture, bulk density and grain hardness (Ratnavathi & Komala, 2016). Endosperm texture is the portion of hard endosperm in relation to the soft endosperm (Rooney & Miller, 1981). Hard (or vitreous) endosperm in sorghum is harder, denser, and more translucent than the soft endosperm, whereas soft endosperm is more porous and opaquer. Endosperm texture is an important factor in the milling quality of grains and resulting flour, as well as susceptibility to grain mold (Menkir et al., 1996; Rooney & Miller, 1981). Generally, kernels with a higher portion of hard endosperm are preferred for milling because they are more resistant to breakage during

decortication and yield cleaner endosperm of larger particle size giving a higher milling yield (Rooney & Miller, 1981). Because endosperm texture is difficult to measure, relatively few studies have examined the genetics controlling this trait (Patil et al., 2019). Therefore, identification of better phenotyping methods may lead to gene discovery, improved selection efficiency, and advances in grain quality.

Computed tomography (CT) imaging technology is a powerful tool that can be utilized to measure complex features in biological specimens. These CT scanners work by beaming x-rays through an object while rotating around the object in a helical path. The resultant x-ray signals are then processed using mathematical algorithms and stitched together into cross sectional images, or "slices," that are stacked together forming a three-dimensional image. From CT images, volumetric data can be analyzed for various structures with different densities. Until recently, the scale, resolution, throughput, accessibility and cost of this technology limited its use (Zhang & Zhang, 2018). However, demand for medical imaging technology has made CT scanning technology accessible to plant scientists.

Utilization of CT imaging offers many potential advantages over traditional phenotyping methods including nondestructive data acquisition, increased throughput and efficiency for gathering multiple traits, and better more accurate measurements (Dhondt et al., 2010; Zhang & Zhang 2018). Multiple vegetative plant structures have been characterized by CT imaging including stems, leaves, and roots and in numerous plant species (Dhondt et al., 2010; Zhu et al., 2012; Gomez et al., 2018; Keszthelyi, Kovács, and Donkó, 2016; Ahmed et al., 2018; Donis-González et al., 2015; Zhang &

Zhang, 2018). In sorghum, Gomez et al. (2018) developed a high throughput phenotyping system for morpho-anatomical stem properties that could be used in a crop improvement program. Therefore, it may be possible to develop similar methodology for analysis of grain although there are different challenges presented when working with the caryopsis.

CT imaging has been used to analyze grain in cereal grain crops. In rice, CT imaging was used to characterize and distinguish high-amylose from wild-type rice (Zhu et al., 2012). In wheat, CT imaging has been used to study damage caused by sprouting and insect infestation (Suresh & Neethirajan, 2015). Similarly in corn, CT imaging was effective to assess damage from insect feeding as well as estimate kernel hardness (Guelpa et al., 2015; Keszthelyi et al., 2016). When measuring kernel hardness in corn, Guelpa et al. (2015) excluded regions not of interest (cavities and germ) which allowed for true determination of volumes and densities of the endosperm. However, this was done by manual removal which is inefficient for practical application. If similar, but improved methodology could be developed for quantifying traits in sorghum grain, it should be possible to assess sorghum endosperm texture on a three-dimensional basis for the first time. In addition, it may be possible to extract information on other traits such as the spatial distribution of endosperm, endosperm hardness, embryo size, kernel size, pericarp thickness, and identification of waxy endosperm.

CT scans produce a vast volume of images that requires efficient methods of data management and analytics. To extract information from CT scans, the images need to be simplified and partitioned into regions of interest. This process, segmentation, assigns a

label to every pixel in an image based on certain common characteristics. From this, quantitative data can be extracted in the form of size and shape of objects in proportion to one another. The simplest approach to segmenting an image is to use thresholds based on pixel intensity to subdivide an image into different regions. However, there are limitations to this approach when sufficient differences in pixel intensity are lacking as is common in real world applications. Guelpa et al. (2015) reported from CT scans of corn kernels that the density of the germ and hard endosperm were very similar and accurate discrimination between the two was not possible using thresholds based on pixel intensity. Other methods of segmentation include looking for acute changes in pixel intensity (edge detection), or changes in texture.

Machine learning approaches offer the potential to combine a collection of feature selection tools for improved image segmentation. The random forest is a classification method combining random uncorrelated decision trees into one prediction model (Breiman 2001). A decision tree is essentially a series of yes/no questions that lead to a predicted class. The trees are trained on different sets of data and use different features to protect each other from their individual errors. In this classification scheme, individually some trees may be wrong, however most trees will be correct. Random forest is a common machine learning based approach to segment images. Applying machine learning based approaches to segmenting CT images may alleviate the challenges put forth by Guelpa et al. (2015) and prevent the use of manual image cleaning.

One of the challenges in implementing machine learning to a broader range of applications is the knowledge gap between scientists versed in machine learning and applied researchers. Trainable Weka Segmentation is a Fiji plugin that combines a collection of machine learning algorithms with a graphical user interface for ease of accessibility and functionality (Arganda-Carreras et al., 2017). This software is freely available and can help bridge the gap between the machine learning and image processing fields.

Limitations in accurate and nondestructive analysis of grain samples can impede improvements in sorghum grain quality. To this end, a phenotyping platform for analysis of sorghum grain using CT imaging was developed. Thereafter, a diversity panel of sorghum was used to validate the effectiveness of CT imaging to measure structural characteristics in sorghum kernels.

3.2. Materials and methods

3.2.1. Plant material

A panel of 19 sorghum inbred lines selected for grain composition diversity was used for analysis (Table 7). These genotypes varied for kernel traits including pericarp color, mesocarp thickness, presence or absence of the testa layer, kernel size, kernel hardness, and endosperm texture. Grain was bulk harvested from ten panicles for each genotype in 2019 at physiological maturity in College Station, Texas and stored in a cold vault at 11 – 13% moisture until scans were conducted.

Table 7. List of plant material and phenotypic kernel characteristics of 19 sorghum inbred lines evaluated in this study.

Genotype	Pericarp Color	Mesocarp Thickness	Testa	Reference
Ajabsido	White	Thick	Yes	(W. L. Rooney, 2004)
B.OK11	White	Thick	No	(Weibel et al., 1984)
B.Tx2928	White	Thick	No	(Rooney 2003)
B.Tx378	Red	Thick	No	(Stephens & Karper, 1965)
B.Tx399	Red	Thick	No	(Stephens & Karper, 1965)
B.Tx642	Yellow	Thick	No	(Rosenow, 2002)
B.TxArg-1	White	Thin	No	(Miller, Domanski, and Giorda 1992)
Dorado	White	Thin	No	(Clara et al., 1986)
FC6601_Spur Feterita	White	Thick	Yes	(Vinall et al., 1936)
ICSV400	White	Thin	No	(Murty et al., 1998)
ICSV745	White	Thin	No	(ICRISAT, 1994)
R.TX2536	White	Thin	No	(Rosenow, unpublished data, 1964)
R.Tx430	White	Thin	No	(Miller 1984)
SC103-12E (IS12170C)	Red	Thin	Yes	(Rosenow, unpublished data, 1970)
SC283 (IS7173C-TAM)	White	Thin	No	(Rosenow, unpublished data, 1972)
Standard Early Hegari (SN106)	White	Thick	Yes	(Swanson & Laude, 1942)
Sureno	White	Thin	No	(Meckenstock et al., 1993)
TAM2566	Red	Thin	Yes	(Johnson et al., 1982)
Texas Blackhull Kafir (SN59)	White	Thick	No	(Stephens & Karper, 1965)

3.2.2. Experimental details

The experimental design was a randomized complete block design with three replications. An experimental unit, compromised of 40 sorghum kernels of a genotype, was placed in a single well in an expanded polystyrene foam microtube storage box. In total, each CT scan contained 21 experimental units constituted by 19 different genotypes. One entry (R.Tx430) was replicated three times to assess the extent of spatial variation within a scan. Three separate CT scans were completed; each scan was considered a replication.

3.2.3. CT scanning and image processing

The CT scans were performed by a North Star Imaging X50 industrial 3D X-ray inspection system. VorteX automated single pass computed tomography scanning technique, which utilizes spiral acquisition and reconstruction with a digital flat-panel detector, was used to construct images. This alleviated the use of volume stitching and provided uniform resolution in axial and sagittal slices across the entire volume. Scans took approximately 2.5 hours to complete and provided a resolution of 20.2 um. After imaging, 2-dimenstional slices from the X-axis were exported as PNG files. From each scan, approximately 1500 images (~40 GB when uncompressed), were extracted. Stacks of two-dimensional slices were imported into Fiji (Schindelin et al., 2012) where they were converted to 8-bit greyscale to reduce file sizes. Stacks of images were then processed using the enhanced contrast feature with saturated pixels set to 0.3% and normalized using the stack histogram. This processing made manual classification easier and normalized the three different scans to a similar range of pixel values.

3.2.4. Image data extraction

A machine learning based plugin in Fiji, Trainable Weka Segmentation 3D (Arganda-Carreras et al., 2017), was used to segment stacks of images into different regions. A classifier was manually trained on a set of images, then applied to the stacks of images to segment them into regions containing background, pericarp, embryo, soft endosperm, or hard endosperm. The classifier was built using a fast random forest, a multithreaded version of the random forest put forth by Breiman (2001), with 200 trees and two random features per node. The tree number was selected as a compromise between

computational power and the diminishing returns in performance gain with more trees (Oshiro et al., 2012; Probst & Boulesteix, 2018). All training features available were evaluated and the best ones were chosen based on out-of-bag error, computational requirements, and visual validation. The selected training features were Hessian, mean, and variance with a minimum sigma of one and a maximum sigma of eight. Mean and variance are texture-based filters useful in differentiating between areas that do not have distinct boundaries but contain patterns of homogenous variation. In the present study, this was useful in delimiting between the regions of sorghum kernels where there was a gradation in pixel values as opposed to distinct boundaries. Hessian is an edge detection filter useful in discriminating the borders of objects defined by clear boundaries such as that between the caryopsis and background. File sizes were too large segment all at once so stacks of images were subdivided into 21 smaller tiles of more manageable size. Every genotype was allocated to an individual tile and segmented separately using the common classifier. Pixel counts for each of the five classes were obtained for every experimental unit and retained for further use.

Total pixel number of kernels for each experimental unit was calculated by adding the number of pixels containing hard endosperm, soft endosperm, pericarp and embryo. A reference point in the image was measured to find pixels per mm. Average single kernel volume was then calculated by converting total pixel number to mm^3 and dividing that by 40 (the numbers of kernels/per entry). Average embryo, endosperm, and pericarp volume were calculated similarly to average kernel volume using the pixel number for each respective region. Endosperm texture was calculated by dividing the

total number of pixels containing hard endosperm, by the total number of pixels containing soft endosperm for each experimental unit.

Not using the segmentation classifier, average endosperm intensity of each experimental unit was calculated by averaging the range of pixel values in all 40 kernels across all slices. Pixel values range in brightness from 1 to 255, where higher pixel values are brighter and represent denser objects in a CT scan. Pixel values below 70 were ignored as background, and pixel values above 248 were ignored as embryo. This was done to achieve an approximation of endosperm hardness based on density, hence exclusion of background noise and regions containing pericarp or embryo.

3.2.5. Ground-truth data collection for validation

Reference grain quality parameters for each genotype were established using both quantitative and visual subjective tests. First, the Single Kernel Characterization System (SKCS) (SKCS 4100, Perten Instruments North America Inc., Springfield, IL) was used to measure diameter, weight, and hardness of 300 individual kernels. This method is widely used in the wheat industry and accepted by the sorghum industry as a tool for measuring grain quality parameters (Bean et al., 2006).

Visual assessment of endosperm texture was estimated by cutting three kernels from each genotype longitudinally along the embryo to bisect the caryopsis, and visually scoring them based on the ratio of hard to soft endosperm. Genotypes were placed into categories from one to five where one is greater than 80% soft endosperm, two is 80% to 60% soft endosperm, three is 60% to 40% soft endosperm, four is 40% to 20% soft endosperm, and five is less than 20% soft endosperm. This was done analogous to

traditional phenotyping methods in which the reliability of data is subject to the skill and expertise of the scorer (Maxson et al., 1971).

3.2.6. Statistical analysis

Restricted maximum likelihood analysis (REML) was conducted in JMP using the model:

$$Y_{ij} = u + Gen_j + Rep_i + Col_k + Row_l + E_{ij}$$

where Y_{ij} is the trait of interest, u is the mean effect, Gen_j is the effect of the jth genotype, Rep_i is the effect of the ith replicate, Col_k is the effect of the kth column, Row_l is the effect of the lth row, and E_{ij} is the random error term. Inclusion of spatial corrections, row and column, was done to assess and account for variance within CT scans. It was hypothesized that things in the center of a CT scan may appear denser possibly due to changes in attenuation from the x-ray passing through more material, or if the imaging gantry does not travel far enough past the ends of the sample to record an accurate image. Factors with negative variance components were removed from the model. Normality of residuals from the models were checked using the Anderson-Darling test, and log base ten transformations were used to normalize non-normal traits. All random models were used to generate estimated best linear unbiased predictors (EBLUPs) for each genotype, variance components, and repeatability estimates. Repeatability (R) on an entry-mean basis was calculated using the equation:

$$R = (\sigma^2_g) / (\sigma^2_g + (\sigma^2_e / rep))$$

where R = the repeatability, σ^2_g = the genotypic variance, σ^2_e = the error variance, and rep = the number of replications. Repeatability, calculated similar to heritability,

indicates the consistency of data and is used in the absence of family structure. Pearson correlation coefficients (r) were computed to assess the relationship between EBLUPs of CT-derived traits and validate against ground-truth data. Best linear unbiased estimators (BLUEs) were estimated using the aforementioned models with genotype being considered as a fixed effect and all other factors random effects. The Tukey – Kramer honestly significance difference (HSD) test was used to determine if genotypes were significantly different from one another using BLUEs.

3.3. Results and discussion

3.3.1. Phenotypic variation

Significant ($p < .01$) genetic variation was detected for embryo volume, endosperm intensity, endosperm texture, endosperm volume, pericarp volume, and kernel volume using CT imaging. Across traits, most of the percent variation was explained by the all random effects model (Figure 3). Residual errors were small which resulted in high R^2 values and repeatability estimates for all traits (Table 8).

Some traits, such as CT Endosperm Intensity, had a large replication effect (Figure 3). This replication effect was likely due to subtle differences in average intensity values between scans (e.g., some scans were brighter than others). Therefore, if data is extracted from multiple CT scans, control genotypes or reference objects of known density are necessary to normalize all scans to the same range of intensity values. Traits derived using the machine learning classifier were less affected than endosperm intensity by replication effects because a combination of features (e.g., mean, variance and Hessian) was used as opposed to pixel intensity alone (Figure 3).

Spatial variation, accounted for in the model by row and column position in CT scans, were not significant sources of variation for any traits but a small amount of variance was partitioned to spatial effects. Since a conservative approach was used to assess significance of effects, spatial variation may still be present within scans. As previously mentioned, spatial variation may be due to due to changes in attenuation from the x-ray passing through more material, or if the imaging gantry did not travel far enough past the ends of the sample to record an accurate image.

Across all genotypes, sorghum kernel composition measured using CT imaging averaged 9% pericarp, 76% endosperm, and 14% embryo (Table 8). Previous literature reported kernel composition as 7% pericarp, 84% endosperm, and 9% embryo (Bean et al., 2016; Bidwell et al., 1922), so CT-estimates are slightly higher for pericarp and embryo with a concomitant reduction in endosperm proportion. Average kernel volume determined by CT imaging was 19.9 mm^3; the most recent U.S. Grains Council survey reported similar US sorghum kernel volumes of 19.34 mm^3 in 2015 and 20.57 mm^3 in 2016 (Bard & Schroeder, 2016).

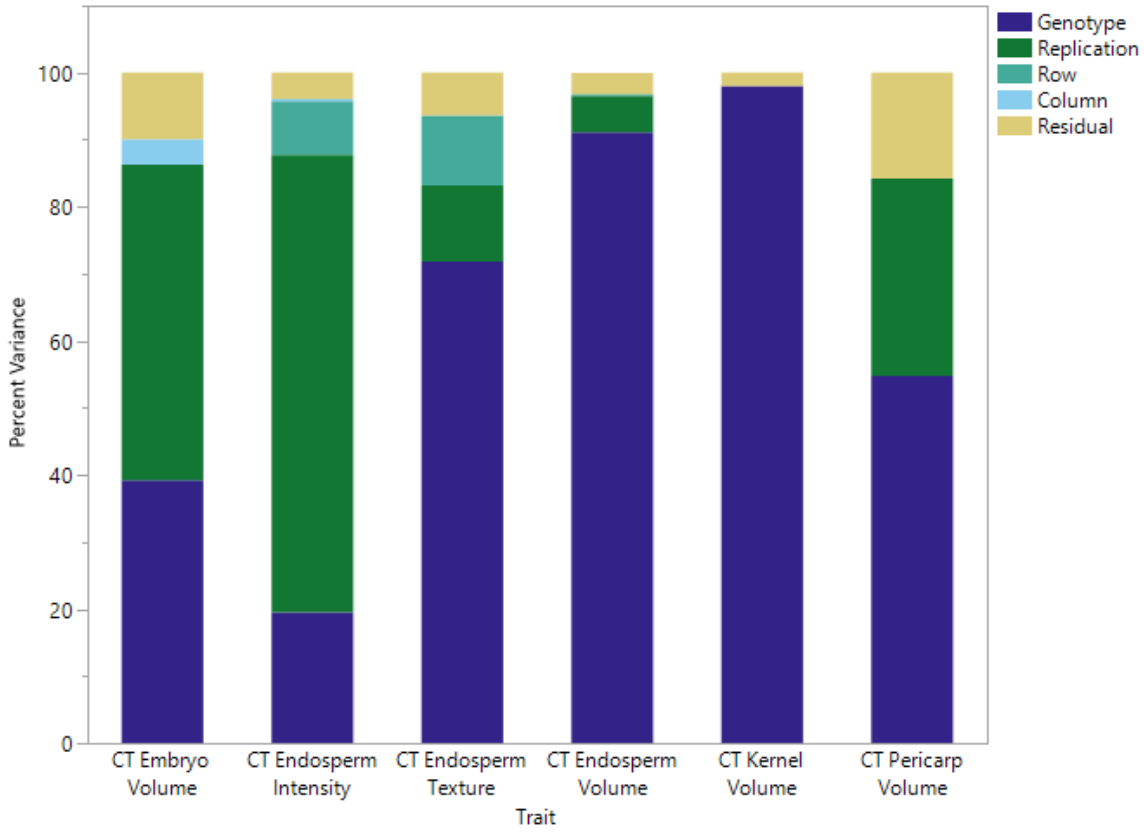

Figure 3. Percent variance associated with factors in CT-derived estimates of sorghum kernel structure for 19 different sorghum genotypes. Replication refers to independent CT scans and while row and column refer to the spatial position within the CT machine.

Table 8. Best linear unbiased estimators for CT-derived measures of sorghum grain structure for 19 sorghum genotypes in which total variation explained by the model (R^2), and repeatability are presented.

Genotype	CT Seed Size (mm^3)	CT Pericarp Volume (mm^3)	CT Embryo Volume (mm^3)	CT Endosperm Volume (mm^3)	CT Endosperm Texture	CT Endosperm Intensity
Ajabsido	29.12	2.21	4.05	22.77	1.79	181.59
B.OK11	14.30	1.54	2.03	10.77	1.43	178.68
B.Tx2928	16.94	1.76	2.53	12.68	2.19	181.27
B.Tx378	20.56	2.03	2.81	15.56	1.69	179.17
B.Tx399	24.13	2.02	3.33	18.79	2.69	185.63
B.TxArg-1	12.55	1.30	1.82	9.24	2.98	183.87
B.Tx642	18.13	1.76	2.46	13.65	3.51	186.49
Dorado	20.86	1.93	2.62	16.20	2.54	184.34
ICSV400	24.79	2.10	3.74	18.84	2.97	187.32
ICSV745	11.82	1.24	1.73	8.82	3.96	184.98
R.TX2536	22.54	2.05	3.07	17.25	2.29	182.79
R.Tx430	27.00	2.26	4.14	20.46	2.31	182.86
SC103-12E	20.19	2.00	2.72	15.10	1.08	174.55
SC283	15.83	1.51	2.91	11.40	3.99	188.32
Spur Feterita (FC6601)	26.91	2.24	3.68	21.06	1.48	178.72
Standard Early Hegari (SN106)	19.40	1.70	2.54	14.91	1.41	176.68
Sureno	14.68	1.49	2.38	10.97	5.11	187.8
TAM2566	20.22	1.82	2.70	15.71	1.43	179.46
Texas Blackhull Kafir (SN59)	17.66	1.70	2.06	13.74	1.35	175.29
Average	19.87	1.82	2.81	15.15	2.43	182.10
HSD	2.15	0.49	1.36	2.39	1.53	6.14
R^2	0.92	0.97	0.95	0.98	0.88	0.99
Repeatability	0.97	0.94	0.97	0.99	0.91	0.99

Note: HSD = honestly significant difference from Tukey-Kramers test.

3.3.2. Correlations and validation

Strong correlations were observed between many CT-derived trait measurements (Figure 4). As expected, CT kernel volume was correlated ($p < .01$) with CT embryo volume, CT endosperm volume, and CT pericarp volume because larger grains are naturally

comprised of greater volumes of embryo, endosperm, and pericarp. CT endosperm intensity was correlated ($p < .01$) with CT endosperm texture, which is logical given the differences in density between soft and hard endosperm (Rooney and Miller 1981).

Estimates of CT-derived traits were also correlated with ground-truth measurements from SKCS and visual scoring. Visual scoring for endosperm texture was highly correlated ($p < .01$) with both CT endosperm intensity and CT endosperm texture (Figure 4). SKCS kernel hardness was also strongly correlated ($p < .01$) with both CT endosperm intensity and CT endosperm texture (Figure 4). SKCS kernel diameter and SKCS kernel weight were both correlated ($p < .01$) with CT kernel volume (Figure 4). The strong correlations reported herein suggest that CT-derived trait measures are reliable.

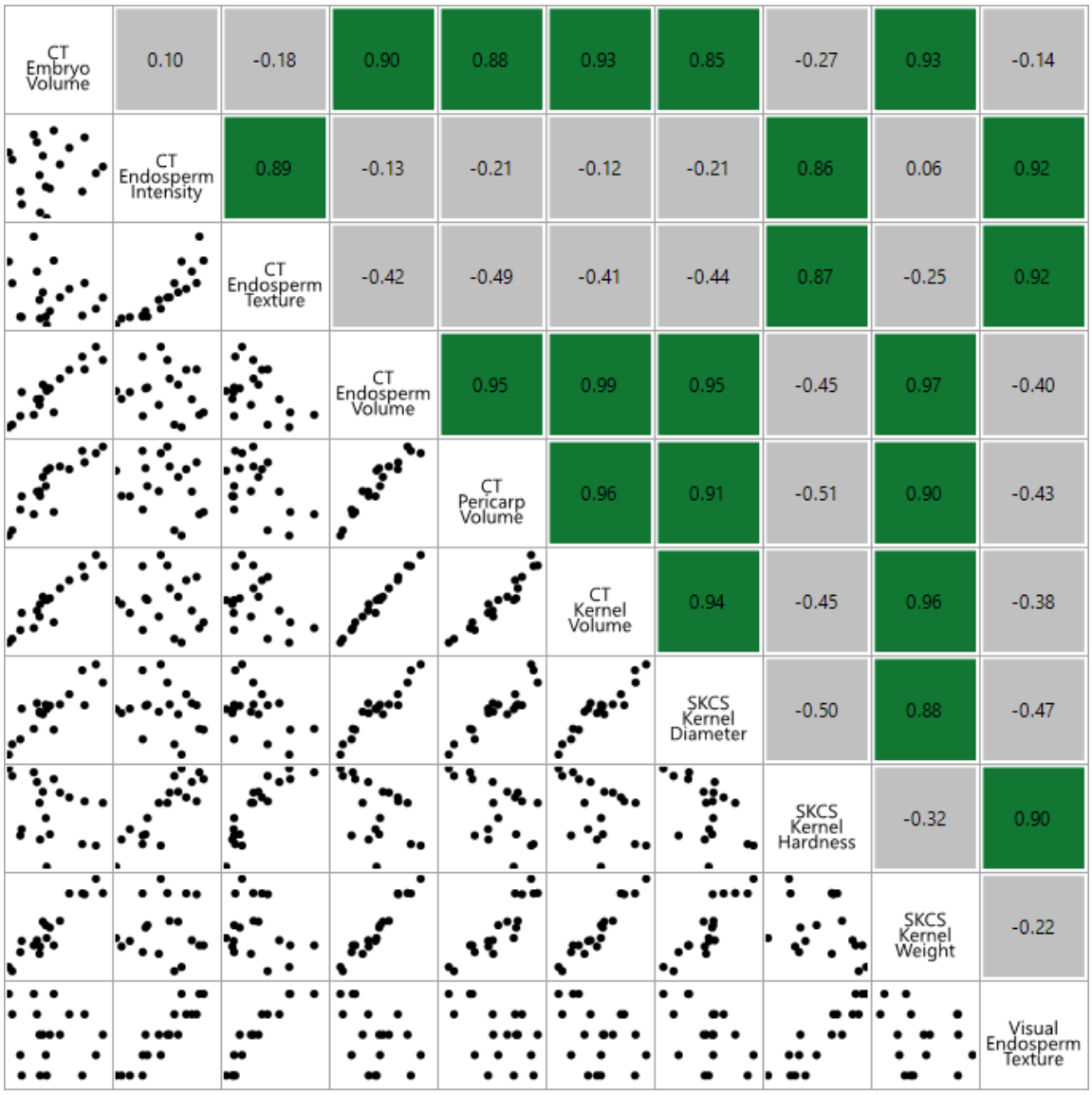

Figure 4. Correlations among different sorghum structural characteristics as measured by CT imaging, SKCS, and visual scoring. Pearson's correlation coefficients significant at $p < .01$ are colored green and shown in the top right. Graphic depictions of the correlation scatterplot matrices are presented in the lower left.

3.3.3. CT imaging challenges

CT images were segmented into regions containing background, embryo, soft endosperm, hard endosperm, and pericarp (Figure 5). In addition, other phenotypic

kernel characteristics, such as the presence of cracks and voids, were observed within some genotypes (Figure 5). Since the segmentation classifier used herein detected these large hollow voids and removed them from kernels, the estimates of sorghum kernel structure remained unbiased.

Histograms of pixel values for scanned grain lacked distinct peaks and valleys corresponding to individual regions of the caryopsis as reported by Guelpa et al. (2015) in corn. This made segmentation more difficult and necessitated the use of a more complex approach than relying singularly on intensity value of regions. Some errors in classification were present, likely due to the lack of marked differences between regions of the kernel. For example, regions of hard endosperm were occasionally misclassified as embryo and regions of the embryo were occasionally misclassified as soft endosperm (Figure 5). This is because even within regions, pixel intensity and texture are not homogenous. For instance, within the embryo, the scutellum is denser and therefore brighter than the embryonic axis. This could lead to higher residual errors in models and either over or underestimation of some regions of the kernel. Potential ways of accounting for this in future analysis involve more robust segmentation classifiers with greater number of features, larger sigma values, and more classes. However, more robust models are computationally more intensive and often impractical as they would slow throughput efficiency of subsequent phenotypic analysis. Overall in this study, misclassifications were minimal and did not negatively impact data quality but they may explain why estimates of embryo and pericarp volume were slightly higher than that reported in previous literature.

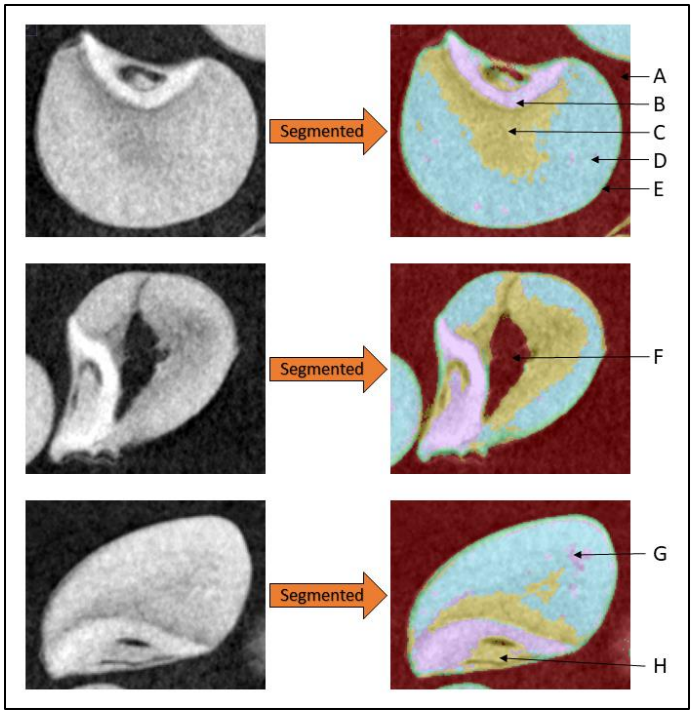

Figure 5. Images of sorghum kernels before and after segmentation. (A) background, (B) embryo, (C) soft endosperm, (D) hard endosperm, (E) pericarp, (F) hollow void in kernel, (G) hard endosperm misclassified as embryo, (H) less dense embryonic axis region of embryo misclassified as soft endosperm.

3.3.4. Genotypic differences

With the methodology presented herein, structural characteristics of sorghum caryopsis can be quantified. This serves many potential applications to plant breeders and cereal chemists alike for use in gene discovery, physiological studies, and other research. One such application is to discriminate between genotypes. Significant genotypic differences were detected between genotypes for all traits using CT imaging (Table 8). For a trait like endosperm texture, CT imaging detected more quantitative genotypic differences with more statistical differences between genotypes than visual scoring (Table 8). In addition, genotypes can be selected for different end-use purposes based on structural

characteristics. For milling, a genotype with larger kernels, higher percent endosperm, and harder endosperm would be preferred such as B.Tx399 (Table 8). Also, cracks and voids were observed in some genotypes, such as Ajabsido, which would be undesirable for milling as kernels would be more prone to breakage during harvest and decortication.

The sorghum lines sampled in this analysis included one genotype with waxy endosperm (B.TxARG-1). Waxy endosperm is caused by a genetic mutation that inhibits the synthesis amylose resulting in a glossy endosperm phenotype that is slightly less dense and phenotypically distinct from normal endosperm (Karper 1933; Rooney and Miller 1981). Efforts to discern between waxy and regular endosperm using CT imaging were not successful. Consequently, the approaches used to measure endosperm properties characterized material similarly regardless of endosperm type. B.TxARG-1 (waxy), was classified as around the same relative ranking for hardness and texture by SKCS, visual, and CT scanning. Therefore, there is no evidence to suggest separate phenotyping methods are needed for waxy and non-waxy genotypes using the phenotyping pipeline provided.

3.4. Conclusions

The phenotyping pipeline presented herein can be automated using the source code in Fiji and did not require manual input past training the initial classifier; thus, increasing throughput efficiency of previously difficult to measure traits. This allowed accurate classification of endosperm texture as well other sorghum kernel structural characteristics. Based on the results presented herein, CT imaging presents new and unique opportunities for scientists to study sorghum grain in a nondestructive manor.

With the capability to three-dimensionally segment sorghum kernels into regions, future studies can assess the spatial distribution and relationship structural characteristics have on other grain quality traits.

4. SUMMARY

Applications of new technology have been at the forefront of innovative plant breeding approaches that help contribute to continual advances in genetic gain. Consider the impact that agrobacterium mediated transformation had on select crop species, or the promise that gene editing, unmanned aerial systems, and genomic prediction models currently present. Adoption of these new technologies is not always a rapid process as it takes time to ascertain the potential benefits and optimize their use. Presented herein, applications to a sorghum breeding program for GBS and CT imaging were evaluated. GBS remains a powerful tool to examine the structure and relationship between germplasm. CT imaging presents new ways to increase throughput and selection efficiency for grain quality related traits in a nondestructive manor.

5. REFERENCES

Ahmed, M. R., Yasmin, J., Collins, W., & Cho, B.-K. (2018). X-ray CT image analysis for morphology of muskmelon seed in relation to germination. *Biosystems Engineering*, *175*, 183–193. https://doi.org/10.1016/j.biosystemseng.2018.09.015

Ahnert, D., Lee, M., Austin, D. F., Livini, C., Woodman, W. L., Openshaw, S. J., Smith, J. S. C., Porter, K., & Dalton, G. (1996). Genetic diversity among elite sorghum inbred lines assessed with DNA markers and pedigree information. *Crop Science*, *36*(5), 1385–1392. https://doi.org/10.2135/cropsci1996.0011183X003600050049x

Amelework, B., Shimelis, H., & Laing, M. (2017). Genetic variation in sorghum as revealed by phenotypic and SSR markers: implications for combining ability and heterosis for grain yield. *Plant Genetic Resources: Characterisation and Utilisation*, *15*(4), 335–347. https://doi.org/10.1017/S1479262115000696

Anderson, W. K., Shackley, B. J., & Sawkins, D. (1998). Grain yield and quality: Does there have to be a trade-off? In H. J. Braun, F. Altay, & W. E. Kronstad (Eds.), *Wheat Prospects for Global Improvment* (pp. 249–254). Springer, Dordrecht. https://doi.org/10.1007/978-94-011-4896-2_34

Arganda-Carreras, I., Kaynig, V., Rueden, C., Eliceiri, K. W., Schindelin, J., Cardona, A., & Sebastian Seung, H. (2017). Trainable Weka Segmentation: a machine learning tool for microscopy pixel classification. *Bioinformatics*, *33*(15), 2424–2426. https://doi.org/10.1093/bioinformatics/btx180

Aruna, C., & Visarada, K. B. R. S. (2019). Other industrial uses of sorghum. In C. Aruna, K. B. R. S. Visarada, B. V. Bhat, & V. A. Tonapi (Eds.), *Breeding Sorghum for Diverse End Uses* (pp. 271–292). Elsevier. https://doi.org/10.1016/b978-0-08-101879-8.00017-6

Axtell, J., Kapran, I., Ibrahim, Y., Ejeta, G., & Andrews, D. J. (1999). Heterosis in sorghum and pearl millet. In J. G. Coors & S. Pandey (Eds.), *The Genetics and Exploitation of Heterosis in Crops* (pp. 375–386). https://doi.org/10.2134/1999.geneticsandexploitation.c35

Bard, S., & Schroeder, C. (2016). *Sorghum harvest quality report.* Washington D. C.: U.S. Grains Council.

Basnet, B. R., Crossa, J., Dreisigacker, S., Pérez-Rodríguez, P., Manes, Y., Singh, R. P., Rosyara, U. R., Camarillo-Castillo, F., & Murua, M. (2019). Hybrid wheat prediction using genomic, pedigree, and environmental covariables interaction models. *Plant Genome*, *12*(1). https://doi.org/10.3835/plantgenome2018.07.0051

Bean, S. R., Chung, O. K., Tuinstra, M. R., Pedersen, J. F., & Erpelding, J. (2006).

Evaluation of the single kernel characterization system (SKCS) for measurement of sorghum grain attributes. *Cereal Chemistry, 83*(1), 108–113. https://doi.org/10.1094/CC-83-0108

Bean, S. R., Wilson, J. D., Moreau, R. A., Galant, A., Awika, J. M., Kaufman, R. C., Adrianos, S. L., Ioerger, B. P., Ciampitti, I., & Prasad, V. (2016). Structure and composition of the sorghum grain. In I. Ciampitti & V. Prasad (Eds.), *Sorghum: State of the Art and Future Perspectives*. ASA, CSSA, and SSSA. https://doi.org/10.2134/agronmonogr58.2014.0081

Bidwell, G. L., Bopst, L. E., & Bowling, J. D. (1922). *A physical and chemical study of milo and feterita kernels*. Washington D. C.: United States Department of Agriculture

Bouis, H. E., & Welch, R. M. (2010). Biofortification—a sustainable agricultural strategy for reducing micronutrient malnutrition in the global south. *Crop Science, 50*, S-20-S-32. https://doi.org/10.2135/cropsci2009.09.0531

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brown, P. J., Myles, S., & Kresovich, S. (2011). Genetic support for phenotype-based racial classification in sorghum. *Crop Science, 51*(1), 224–230. https://doi.org/10.2135/cropsci2010.03.0179

Burrell, M., Sharma, A., Patil, N. Y., Collins, S. D., Anderson, W. F., Rooney, W. L., & Klein, P. E. (2015). Sequencing of an anthracnose-resistant sorghum genotype and mapping of a major QTL reveal strong candidate genes for anthracnose resistance. *Crop Science, 55*(2), 790–799. https://doi.org/10.2135/cropsci2014.06.0430

Clara, R. V., Cordova, H. S., & Coto, H. A. (1986). *Siembra maicillo (sorgo) ISIAP dorado*. La Libertad: CENTA

Dhondt, S., Vanhaeren, H., Van Loo, D., Cnudde, V., & Inzé, D. (2010). Plant structure visualization by high-resolution X-ray computed tomography. *Trends in Plant Science, 15*(8), 419–422. https://doi.org/10.1016/j.tplants.2010.05.002

Donis-González, I. R., Guyer, D. E., Chen, R., & Pease, A. (2015). Evaluation of undesirable fibrous tissue in processing carrots using computed tomography (CT) and structural fiber biochemistry. *Journal of Food Engineering, 153*, 108–116. https://doi.org/10.1016/j.jfoodeng.2014.12.012

Duvick, D. N. (2005). The contribution of Breeding to Yield advances in maize (*Zea mays* L.). *Advances in Agronomy, 86*, 83–145. https://doi.org/10.1016/S0065-2113(05)86002-X

Duvick, D. N. (1999). Heterosis: feeding people and protecting natural resources. In J. G. Coors & S. Pandey (Eds.), *The Genetics and Exploitation of Heterosis in Crops* (pp. 19–29). https://doi.org/10.2134/1999.geneticsandexploitation.c3

Earp, C. F., & Rooney, L. W. (1982). Scanning electron microscopy of the pericarp and testa of several sorghum varieties. *Food Structure, 1*(2), 125–134.

Flint-Garcia, S. A., Buckler, E. S., Tiffin, P., Ersoz, E., & Springer, N. M. (2009). Heterosis is prevalent for multiple traits in diverse maize germplasm. *PLoS ONE, 4*(10). https://doi.org/10.1371/journal.pone.0007433

Gizzi, G., & Gambin, B. L. (2016). Eco-physiological changes in sorghum hybrids released in Argentina over the last 30 years. *Field Crops Research, 188*, 41–49. https://doi.org/10.1016/j.fcr.2016.01.010

Glueck, J. A., & Rooney, L. W. (1980). Chemistry and structure of grain in relation to mold resistance. In G. D. Bengtson (Ed.), *International Workshop on Sorghum Diseases* (pp. 119–140). Hyderabad: ICRISAT.

Gomez, F. E., Carvalho, G., Shi, F., Muliana, A. H., & Rooney, W. L. (2018). High throughput phenotyping of morpho-anatomical stem properties using X-ray computed tomography in sorghum. *Plant Methods*, *14*. https://doi.org/10.1186/s13007-018-0326-3

Guelpa, A., Du Plessis, A., Kidd, M., & Manley, M. (2015). Non-destructive estimation of maize (*Zea mays* L.) kernel hardness by means of an X-ray micro-computed tomography (µCT) density calibration. *Food and Bioprocess Technology, 8*(7), 1419–1429. https://doi.org/10.1007/s11947-015-1502-3

Hallauer, A. R. (1999). Temperate maize and heterosis. In J. G. Coors & S. Pandey (Eds.), *The Genetics and Exploitation of Heterosis in Crops* (pp. 353–361). https://doi.org/10.2134/1999.geneticsandexploitation.c33

Horne, D. W. (2019). *Implementation of genomic and phenomic tools for introgression of reinstated sorghum conversion (RSC) germplasm* (Doctoral dissertation, Texas A&M University).

Hubbard, J. E., Hall, H. H., & Earle, F. R. (1950). Composition of the component parts ot the sorghum kernel. *Cereal Chemisrty, 27*, 415–420.

ICRISAT. (1994). *Midge-resistant sorghum cultivar ICSV 745*. Andhra Pradesh: ICRISAT

Johnson, J. W., Rosenow, D. T., Teetes, G. L., & Phillips, J. M. (1982). Registration of 19 greenburg resistant sorghum germplasm lines. *Crop Science, 22*(6), 1272.

https://doi.org/10.2135/cropsci1982.0011183x002200060075x

Jordan, D. R., Tao, Y., Godwin, I. D., Henzell, R. G., Cooper, M., & McIntyre, C. L. (2003). Prediction of hybrid performance in grain sorghum using RFLP markers. *Theoretical and Applied Genetics*, *106*(3), 559–567. https://doi.org/10.1007/s00122-002-1144-5

Karper, R. E. (1933). Inheritance of waxy endosperm in sorghum. *The Journal of Heredity*, *24*(6), 257–262.

Keszthelyi, S., Kovács, G., & Donkó, T. (2016). Computer tomography-assisted imaging analysis in damaged maize grain caused by sitotroga cerealella. *Journal of Plant Diseases and Protection*, *123*(2), 89–92. https://doi.org/10.1007/s41348-016-0009-0

Klein, R. R., Miller, F. R., Dugas, D. V., Brown, P. J., Burrell, A. M., & Klein, P. E. (2015). Allelic variants in the PRR37 gene and the human-mediated dispersal and diversification of sorghum. *Theoretical and Applied Genetics*, *128*(9), 1669–1683. https://doi.org/10.1007/s00122-015-2523-z

Liu, K., & Muse, S. V. (2005). PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics*, *21*(9), 2128–2129. https://doi.org/10.1093/bioinformatics/bti282

Meckenstock, D. H., Gomez, F., Rosenow, D. T., & Guiragossian, V. (1993). Registration of 'Sureño' sorghum. *Crop Science*, *33*(1), 213–213. https://doi.org/10.2135/cropsci1993.0011183x003300010058x

Menkir, A., Ejeta, I. G., Butler, L., & Melakeberhan, A. (1996). Physical and chemical kernel properties associated with resistance to grain mold in sorghum. *Cereal Chemisrty*, *73*(5), 613–617.

Menz, M. A., Klein, R. R., Unruh, N. C., Rooney, W. L., Klein, P. E., & Mullet, J. E. (2004). Genetic diversity of public inbreds of sorghum determined by mapped AFLP and SSR markers. *Crop Science*, *44*(4), 1236–1244. https://doi.org/10.2135/cropsci2004.1236

Miller, F. R. (1984). Registration of RTx430 sorghum parental line. *Crop Science*, *24*(6), 1224–1224. https://doi.org/10.2135/cropsci1984.0011183x002400060074x

Miller, Fred R., Domanski, C., & Giorda, L. M. (1992). Registration of A/BTxARG-1 sorghum. *Crop Science*, *32*(6), 1517. https://doi.org/10.2135/cropsci1992.0011183x003200060057x

Mindaye, T. T., Mace, E. S., Godwin, I. D., & Jordan, D. R. (2015). Genetic

differentiation analysis for the identification of complementary parental pools for sorghum hybrid breeding in Ethiopia. *Theoretical and Applied Genetics*, *128*(9), 1765–1775. https://doi.org/10.1007/s00122-015-2545-6

Mindaye, T. T., Mace, E. S., Godwin, I. D., & Jordan, D. R. (2016). Heterosis in locally adapted sorghum genotypes and potential of hybrids for increased productivity in contrasting environments in Ethiopia. *Crop Journal*, *4*(6), 479–489. https://doi.org/10.1016/j.cj.2016.06.020

Moll, R. H., Lonnquist, J. H., Fortuno, E. Z., & Johnson, E. C. (1965). The relationship of heterosis and genetic divergence in maize. *Genetics*, *52*(1), 139–144.

Morishige, D. T., Klein, P. E., Hilley, J. L., Sahraeian, S. M. E., Sharma, A., & Mullet, J. E. (2013). Digital genotyping of sorghum-a diverse plant species with a large repeat-rich genome. *BMC Genomics*, *14*(1). https://doi.org/10.1186/1471-2164-14-448

Murty, D. S., Nwasike, C. C., & Bello, S. A. (1998). Registration of 'ICSV 400' sorghum cultivar. *Crop Science*, *38*(6), 1717–1718. https://doi.org/10.2135/cropsci1998.0011183x003800060060x

Nie, Y., Ji, W., & Ma, S. (2019). Assessment of heterosis based on genetic distance estimated using SNP in common wheat. *Agronomy*, *9*(2). https://doi.org/10.3390/agronomy9020066

Niebur, W. S., Rafalski, A. J., Smith, O. S., & Cooper, M. (2004). Applications of genomics technologies to enhance rate of genetic progress for yield of maize within a commercial breeding program. In Fischer T. et al. (Eds.), *Proceedings of the 4th International Crop Science Congress, Brisbane, Australia: New Directions for a Diverse Planet*. Retrieved from http://www.cropscience.org.au

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). How Many Trees in a Random Forest? In P. Perner (Ed.), *International workshop on machine learning and data mining in pattern recognition* (pp. 154–168). https://doi.org/10.1007/978-3-642-31537-4_13

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., Schmutz, J., Spannagl, M., Tang, H., Wang, X., Wicker, T., Bharti, A. K., Chapman, J., Feltus, F. A., Gowik, U., … Rokhsar, D. S. (2009). The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, *457*(7229), 551–556. https://doi.org/10.1038/nature07723

Patil, N. Y., Pugh, N. A., Klein, R. R., Martinez, H. S., Martinez, R. S., Rodriguez-Herrera, R., Rooney, W. L., & Klein, P. E. (2019). Heritability and quantitative trait loci of composition and structural characteristics in sorghum grain. *Journal of Crop*

Improvement, *33*(1), 1–24. https://doi.org/10.1080/15427528.2018.1536006

Pfeiffer, B. K., Pietsch, D., Schnell, R. W., & Rooney, W. L. (2019). Long-term selection in hybrid sorghum breeding programs. *Crop Science*, *59*(1), 150–164. https://doi.org/10.2135/cropsci2018.05.0345

Powell, W., Morgante, M., Andre, C., Hanafey, M., Vogel, J., Tingey, S., & Rafalski, A. J. (1996). The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. *Molecular Breeding*, *2*(3), 225–238. https://doi.org/10.1007/BF00564200

Probst, P., & Boulesteix, A.-L. (2018). To tune or not to tune the number of trees in random forest. *Journal of Machine Learning Research*, *18*, 1–18.

Quinby, J. R. (1963). Manifestations of hybrid vigor in sorghum. *Crop Science*, *3*(4), 288. https://doi.org/10.2135/cropsci1963.0011183x000300040003x

Ramu, P., Billot, C., Rami, J. F., Senthilvel, S., Upadhyaya, H. D., Ananda Reddy, L., & Hash, C. T. (2013). Assessment of genetic diversity in the sorghum reference set using EST-SSR markers. *Theoretical and Applied Genetics*, *126*(8), 2051–2064. https://doi.org/10.1007/s00122-013-2117-6

Ratnavathi, C. V., & Komala, V. V. (2016). Sorghum grain quality. In C. V. Ratnavathi, J. V. Patil, & U. D. Chavan (Eds.), *Sorghum Biochemistry: An Industrial Perspective* (pp. 1–61). Elsevier Inc. https://doi.org/10.1016/B978-0-12-803157-5.00001-0

Rooney, Loyd W., & Miller, F. R. (1981). Variation in the structure and kernel characteristics of sorghum. In L. W. Rooney & D. S. Murty (Eds.), *International Symposium on Sorghum Grain Quality*, (pp. 143–162). Patancheru, India: ICRISAT

Rooney, W. L. (2003). Registration of Tx2921 through Tx2928 sorghum germplasm lines. *Crop Science*, *43*(1), 443. https://doi.org/10.2135/cropsci2003.4430

Rooney, W. L. (2004). Sorghum improvment-integrating traditional and new technology to produce improved genotypes. *Advances in Agronomy*, *83*, 37–109. https://doi.org/10.1016/S0065-2113(04)83002-5

Rosenow, D. T. (2002). Release proposal for four A/B sorghum parental lines. Lubbock: Texas A&M Agrilife Research and Extension

Scheet, P., & Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics*, *78*(4), 629–644. https://doi.org/10.1086/502802

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J. Y., White, D. J., Hartenstein, V., Eliceiri, K., Tomancak, P., & Cardona, A. (2012). Fiji: An open-source platform for biological-image analysis. *Nature Methods*, *9*(7), 676–682. https://doi.org/10.1038/nmeth.2019

Schnable, P. S., & Springer, N. M. (2013). Progress toward understanding heterosis in crop plants. *Annual Review of Plant Biology*, *64*(1), 71–88. https://doi.org/10.1146/annurev-arplant-042110-103827

Schnell, R. W. (2018). *Hybrid selection – plant height and grain yield*. College Station: Texas A&M Agrilife Extension - Sorghum Tips.

Shull, G. H. (1908). The composition of a field of maize. *Journal of Heredity*, *os-4*(1), 296–301. https://doi.org/10.1093/jhered/os-4.1.296

Shull, G. H. (1914). Duplicate genes for capsule-form in *Bursa bursa-pastoris*. *Zeitschrift Für Induktive Abstammungs- Und Vererbungslehre*, *12*(1), 97–149. https://doi.org/10.1007/BF01837282

Stephens, J. C., & Holland, R. F. (1954). Cytoplasmic male-sterility for hybrid sorghum seed production. *Agronomy Journal*, *46*(1), 20–24. https://doi.org/10.2134/agronj1954.00021962004600010006x

Stephens, J. C., & Karper, R. E. (1965). *Release of breeding stocks of male-sterilized grain sorghum lines*. College Station: Texas A&M University

Stupar, R. M., Gardiner, J. M., Oldre, A. G., Haun, W. J., Chandler, V. L., & Springer, N. M. (2008). Gene expression analyses in maize inbreds and hybrids with varying levels of heterosis. *BMC Plant Biology*, *8*. https://doi.org/10.1186/1471-2229-8-33

Suresh, A., & Neethirajan, S. (2015). Real-time 3D visualization and quantitative analysis of internal structure of wheat kernels. *Journal of Cereal Science*, *63*, 81–87. https://doi.org/10.1016/j.jcs.2015.03.006

Swanson, A. F., & Laude, H. H. (1942). Sorghums for Kansas. In *Bulletin* (Vol. 304). Manhattan: Kansas State University

Thornton, P. K. (2010). Livestock production: Recent trends, future prospects. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1554), 2853–2867. https://doi.org/10.1098/rstb.2010.0134

Vinall, H. N., Stephens, J. C., & Martin, J. H. (1936). Identification, history and distribution of common sorghum varieties. In *Technical Bulletin* (Issue 506). Washington D. C.: United States Department of Agriculture

Voss-Fels, K. P., Cooper, M., & Hayes, B. J. (2019). Accelerating crop genetic gains with genomic selection. *Theoretical and Applied Genetics*, *132*(3), 669–686. https://doi.org/10.1007/s00122-018-3270-8

Waniska, R. D., & Rooney, L. W. (2001). Structure and chemistry of the sorghum caryopsis. In C. W. Smith & R. A. Frederiksen (Eds.), *Origin, History, Technology and Production* (pp. 649–688). Wiley, New York.

Wegary, D., Vivek, B., & Labuschagne, M. (2013). Association of parental genetic distance with heterosis and specific combining ability in quality protein maize. *Euphytica*, *191*(2), 205–216. https://doi.org/10.1007/s10681-012-0757-2

Wei, X., & Zhang, J. (2018). The optimal mating distance resulting from heterosis and genetic incompatibility. *Science Advances*, *4*(11). https://doi.org/10.1126/sciadv.aau5518

Weibel, D. E., Seiglinger, J. B., & Davies, F. F. (1984). Registration of fourteen sorghum parental lines. *Crop Science*, *24*(3), 628–628. https://doi.org/10.2135/cropsci1984.0011183x002400030064x

Wu, J. W., Hu, C. Y., Shahid, M. Q., Guo, H. Bin, Zeng, Y. X., Liu, X. D., & Lu, Y. G. (2013). Analysis on genetic diversification and heterosis in autotetraploid rice. *SpringerPlus*, *2*(1), 1–12. https://doi.org/10.1186/2193-1801-2-439

Zhang, T., Xian-Lin, N. I., Jiang, K.-F., Deng, H.-F., Qian-Hua, Y., Li, Y., Xian-Qi, W., Cao, Y.-J., & Jia-Kui, Z. (2010). Relationship between heterosis and parental genetic distance based on molecular markers for functional genes related to yield traits in rice. *Rice Science*, *17*(4), 288–295. https://doi.org/10.1016/S1672-6308(09)60029-9

Zhang, Y., & Zhang, N. (2018). Imaging technologies for plant high-throughput phenotyping: a review. *Frontiers of Agricultural Science and Engineering*, *5*(4), 406–419. https://doi.org/10.15302/J-FASE-2018242

Zhu, L.-J., Dogan, H., Gajula, H., Gu, M.-H., Liu, Q.-Q., & Shi, Y.-C. (2012). Study of kernel structure of high-amylose and wild-type rice by X-ray microtomography and SEM. *Journal of Cereal Science*, *55*(1), 1–5. https://doi.org/10.1016/j.jcs.2011.08.013