Impact of Sorghum Racial Structure and Diversity on Genomic Prediction of Grain Yield Components

Sirjan Sapkota,* Richard Boyles, Elizabeth Cooper, Zachary Brenton, Matthew Myers, and Stephen Kresovich

Affiliations: S. Sapkota, Z. Brenton, M. Myers and S. Kresovich, Advanced Plant Technology Program, Clemson University, Clemson, SC 29634; S. Sapkota, R. Boyles, and S. Kresovich, Department of Plant and Environmental Sciences, Clemson University, Clemson SC 29634; R. Boyles, Pee Dee Research and Education Center, Clemson University, Florence, SC 29506; E. Cooper, Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, NC 28223. Received _____. *Corresponding author (ssapkot@g.clemson.edu).

Abbreviations: AR, across race; BL, terminal branch length; BLUE, best linear unbiased estimator; CV, cross validation; DTA, days to anthesis; FLH, flag leaf height; GEBV, genomic estimated breeding values; GN, grain number; GW, grain weight; GY, grain yield; LD, linkage disequilibrium; MAF, minor allele frequency; PCA, principal component analysis; PH, plant height; PL, panicle length; GBLUP: genomic best linear unbiased prediction; *r*: prediction accuracy; SRT, single race training; WR, within race.

ABSTRACT

Population structure is an important factor that affects the accuracy of estimated breeding values in genomic prediction. Natural sorghum populations exhibit population structure resulting from genetic and morphological differentiation due to evolutionary divergence. To study the impact of sorghum racial structure and diversity in genomic prediction, we conducted two cross validation (CV) experiments: CV1; proportional sampling from races, and 2) CV2; sampling from across race (AR) or within race (WR). A diversity panel with 389 individuals with 224,007 single nucleotide polymorphisms were used for genomic prediction. Genomic heritabilities for traits were positively correlated (0.63) with their mean prediction accuracy (*r)* from CV1, and within-subpopulation variance accounted for about 80% of total genetic variance. CV1 prediction accuracy ranged from 0.52 to 0.69, but *r* declined by 39% and 54% on average for WR and AR methods, respectively. As a predictor race explained 30 to 50% of covariance for grain and panicle traits but race was a bad predictor of plant height, as expected. Grain weight was consistently the best predicted trait across CV1 and CV2 methods except in AR. Difference in average *r* for WR and AR was greater in durra and caudatum, small in kafir, and non-existent in guinea and mixed. We observed higher prevalence of minor alleles among guinea and mixed subgroups highlighting contribution of allelic diversity towards prediction accuracy. Genomic prediction in sorghum will benefit from utilization of inter-racial diversity and we emphasize the need for further investigations into the role of racial structure in genomic prediction.

INTRODUCTION

Cultivated crops have undergone genetic bottlenecks as a result of domestication and artificial selection. Genetic diversity in modern crops is further reduced by current plant breeding practices because most of the cultivars are derived from genetically-related varieties that represent a very small fraction of the global diversity of plant germplasm for any species (McCouch, et al., 2013). Effective utilization of genetic diversity to increase resilience and crop yield potential will remain a key aspect to meet the projected food demands in the next few decades and reduce vulnerability to biotic and abiotic stresses,.

Sorghum is an important cereal crop grown and consumed as a staple by over half a billion people in the semi-arid tropics. The earliest known record of sorghum seeds are the charred remains from 8000 years before present found at Nabta Playa near the Egyptian-Sudanese border during archeological excavations (Wendorf, et al., 1992). After early domestication likely near the Sahel region of sub-Saharan Africa, further migration and adaptation of early sorghum domesticates occurred across Africa and Asia. Those demographic events led to the evolution of morphologically and geographically diverse groups that are classified into five major races and 10 intermediate races of sorghum (Harlan and de Wet, 1972, Harlan and Stemler, 1976). This phenotype-based classification of sorghum races has been supported by genetic evidences in a global diversity panel (Brown, et al., 2011). Furthermore, the linkage disequilibrium (LD) in sorghum has shown presence of strong genetic bottleneck and patterns of disruptive selection across the sorghum genome as a result of domestication, adaptation, and diversifying selection (Mace, et al., 2013, Morris, et al., 2013, Wang, et al., 2013).

The sorghum conversion program (SCP) was initiated by the United States Department of Agriculture (USDA) in cooperation with Texas A&M University to introduce novel genetic variation from exotic tropical germplasm by converting selected tropical genotypes to temperate adapted, photoperiod-insensitive lines with short stature (Stephens et al. 1967). This ongoing initiative has been the staple source of germplasm for several public and private breeding programs in temperate regions. However, the conversion of tropical lines through repeated backcrossing is an expensive and labor-intensive process. Therefore, only 1000-1500 tropical lines have been converted, and these converted lines represent a limited fraction of USDA and worldwide collection of sorghum germplasm (Bob Klein, personal communication). Recent advances in genomic and computational capabilities present opportunities for identifying effective strategies for introducing and screening of germplasm for novel genetic variation, which can benefit breeders by making selection of prebreeding germplasm more accurate and meaningful.

Genomic prediction (also known as genomic selection or genome-wide selection) is a method to simultaneously estimate the effects of all genetic markers and use those marker effects to estimate breeding values (Meuwissen et al. 2001; Bernardo and Yu 2007). The marker effects are estimated using both genotypic and phenotypic data from a training population, which can then be used to predict genomic estimated breeding value (GEBV) using only the genotypic information in a testing population. The accuracy of prediction is measured as the correlation between GEBVs and true genetic values, often represented by observed phenotypic values. Genomic prediction is usually applied in breeding populations where the training and testing population have a shared pedigree, but its application can be extended to screening and selection for pre-breeding or population improvement (Yu et al. 2016; Gaynor et al. 2017). Every year, an

increasingly larger number of association studies are conducted for allele mining by breeding and genetics programs across the world. Large-volume phenotypic datasets generated from these studies can be applied in the investigation and application of genomic prediction models across ranges of crops and traits. These resources can then be utilized in careful strategies to tap into the large number of gene bank accessions by screening for useful genetic variation with potential to enhance genetic gain (Yu et al. 2016).

The implementation of genomic prediction in a diverse and stratified population, however, requires careful consideration of the genomic relationship and population structure (Jannink et al. 2010; Habier et al. 2007). Population structure has been shown to affect the accuracy of genomic prediction in stratified populations across several crop species (Guo, et al., 2014, Ly, et al., 2013, Norman, et al., 2018). Population structure analysis can be done using non-model based approaches like principal component analysis (Patterson et al. 2006; Price et al. 2006) or model based clustering approaches like ADMIXTURE (Alexander et al. 2009). Incorporating population structure estimates into both association and prediction studies has proven useful, but the methods for including population structure in the model can vary. While the inclusion of population structure as a covariate has been successfully applied in mixed models for association studies, the use of population structure as fixed effects in genomic best linear unbiased prediction (GBLUP) models would be concerning because they already enter into the model as random effects (de los Campos and Sorensen, 2014, Janss, et al., 2012, Price, et al., 2010). One approach to account for population structure in genomic prediction is designing a cross validation scheme that ensures equal representation of each subpopulation in training and validation sets (Albrecht, et al., 2011, Guo, et al., 2014). Alternatively, in order to avoid biased estimation due to the presence of genetic structure and familial relatedness, prediction analysis

107 can be performed by partitioning the genomic variability into within and across group
108 components (Guo, et al., 2014, Norman, et al., 2018, Technow, et al., 2012). Because of distinct
109 racial structure in sorghum, an approach to account for contribution of racial structure in
110 prediction accuracy by decomposing variance-covariance components into expectations due to
111 race and covariance from individuals within a race could be beneficial.

112 A recent simulation study has highlighted the advantages of genomics-assisted recurrent
113 selection over phenotypic recurrent selection in a nascent and small sorghum breeding program,
114 emphasizing the need for further investigations on genomic selection in sorghum (Muleta, et al.,
115 2019). Since inter-racial diversity is important for heterotic gain, application of genomic
116 prediction in sorghum breeding will benefit from investigations into the effect of racial structure
117 on prediction accuracy. While the effect of population structure in genomic prediction has been
118 extensively studied in major cereal crops, the distinctive evolutionary history and racial structure
119 of sorghum merits the need for investigation into the effect of sorghum racial structure in
120 genomic prediction (Guo, et al., 2014, Isidro, et al., 2015, Norman, et al., 2018). A previous
121 study examined the effect of genetic relatedness on genomic prediction of pedigreed male lines
122 in a sorghum breeding program, however, there has been no studies on the role of sorghum racial
123 structure in genomic prediction (Hunt, et al., 2018). The objectives of our study were to estimate
124 genetic structure and diversity among sorghum races and implement genomic prediction for plant
125 architecture and grain yield traits to examine the effect of racial structure on prediction accuracy
126 using a grain diversity panel.

MATERIALS AND METHODS

Plant materials, field design, and phenotyping

The sorghum diversity panel used in this study consisted of 389 diverse sorghum accessions, including 332 accessions from the United States sorghum association panel (SAP) developed by Casa, et al. (2008). Additional accessions were included for diversity and elite grain characteristics (Boyles, et al., 2016). The population was planted in randomized complete block design with two replications in 2013, 2014, and 2017 field seasons at the Clemson University Pee Dee Research and Education Center in Florence, South Carolina. The accessions were assigned to blocks within the replication based on height, maturity, and photoperiod sensitivity. Each plot was two 6.1 m rows spaced 0.726 m apart with a targeted planting density of 130,000 plants ha^{-1} assuming 75% plant establishment rate. In 2017, an average plant density of ~62,350 plants ha^{-1} was calculated based on plant stand count and row length at 24 days after planting (DAP). Fields were irrigated when plants showed signs of stress in order to avoid confounding effects of maturity and varying degree of drought tolerance in the population on yield. The details on agronomic practices for 2013 and 2014 can be found in Boyles, et al. (2016). In 2017, a lay-by of 93 Kg ha^{-1} N was applied at 35 DAP in addition to variable rate of fertilizer (N, P, K) applications before planting. Preemergence and postemergence herbicide applications in 2017 were consistent with 2013 and 2014. A single application of 0.5 L ha^{-1} of Sivanto™ Prime (Bayer CropScience) was administered at 60 DAP to control sugarcane aphid population.

Three consecutive plants from the odd row of each plot was selected for phenotyping in order to prevent biases due to row effect. We also avoided plants from beginning and end of the row to account for border effect. The detail procedures for phenotyping of agronomic and grain phenotypes has previously been described in Boyles et. al. (2016; 2017) . Days to anthesis

(DTA) for each plot was measured as the number of days from planting to when 50% of the plants in the plot were at mid-bloom. Plant height (PH) was measured from ground to the apex of the primary panicle at physiological maturity. Flag leaf height (FLH), panicle length (PL), and terminal branch length (BL) of each plant harvested in 2017 were measured. Flag leaf height was measured as the height from ground to the flag leaf of the plant. Panicle length was measured as length of primary panicle from the terminal branch to the apex of the panicle, and BL was measured as length of the two terminal primary branches, respectively. A more detailed description of these inflorescence architecture traits can be found in Brown, et al. (2006). Grain yield components were phenotyped from primary panicle of three consecutive plants harvested at physiological maturity as previously mentioned. Panicles were air dried to a constant moisture (10-12%) before threshing. Threshed seed were processed through seed counters (Old Mill Model 900-2) to measure grain number per primary panicle (GN). Grain yield per primary panicle (GY) in grams (g) was measured with a Discovery series scale (Ohaus). Subsequently, thousand grain weight (GW) was calculated from GY and GN; GW (g) = (GY/GN) × 1000.

Phenotypic analysis

R statistical software was used for phenotypic analysis (R Development Core Team, 2016). Simple mean of phenotypic values were calculate for each replication with the years. The phenotypic means of the traits were fitted into a linear mixed model analysis using lme4 package in R (Bates, et al., 2015). We fit the following mixed model equation:

$$y_{ijk} = \mu + G_i + E_j + G_iE_j + R_k(E_j) + e_{ijk}, \qquad (1)$$

where y_{ijk} is the phenotypic value for genotype i, year j, and replication k within the year j; μ is the population mean; G_i is the fixed effects of i^{th} genotype; E_j, GE_{ij}, $R_k(E_j)$ are random effects of year, genotype × year, and replication within the year, respectively; and e_{ijk} is the

random effect of residuals, with $e \sim N(0, \sigma_e^2)$. Since phenotypes for PL, BL and FLH were only available from the year 2017, the model was fit with just the random effect of replications within the year and fixed effect of genotypes. Best linear unbiased estimates (BLUEs) for the traits were calculated from the fixed effect of the genotypes. Correlation plots and histograms for the estimated phenotypic means were generated using the *pairs.panels* function within the R package Psych (Revelle, 2011).

Genotyping

Genetic characterization of the diversity panel was done using genotyping-by-sequencing (GBS) as previously described in Morris, et al. (2013). Sequenced reads were aligned to the sorghum reference assembly (BTx623 v3.1, www.phytozome.net) using burrow-wheelers aligner (Li and Durbin 2010). SNP calling, imputation and filtering were done using the TASSEL 5.0 pipeline (Glaubitz, et al., 2014). A total of 515,318 SNPs was called and subsequently imputed using the *FILLINFindHaplotypesPlugin* and *FILLINImputationPlugin* in TASSEL. FILLIN (Fast, Inbred Line Library ImputatioN) imputes missing genotypes by: (1) haplotype generation using inbred segments that share identity by state, and (2) imputation of resulting haplotypes back onto the target samples (Swarts, et al., 2014). SNP sites were filtered to remove sites with a minor allele frequency (MAF) < 0.01, and sites present in at least 90% of the individuals were retained. A final SNP matrix with a total of 224,007 SNPs was created and used for subsequent genomic analysis and predictions. In the final SNP matrix, all genotypes had less than 10% missing sites. The final SNP genotype matrix was further filtered to retain SNPs with MAF > 0.05 for estimation of the decay of linkage disequilibrium (LD).

Population structure and genetic diversity

The final SNP matrix was used to first identify the optimum number of clusters based on cross validation of error, then used to calculate the ancestry coefficients using ADMIXTURE (Alexander, et al., 2009). Admixture ancestry coefficients (Q matrix) were estimated using the default block relaxation algorithm. We also calculated covariances for the first five principal components (PCs) using the SNP data in TASSEL. A common subset of 35,277 SNPs between the diversity panel and *S. propinquum* from Mace, et al. (2013) was used for neighbor joining tree estimation using the *Cladogram* function in TASSEL. This function first calculates distance between each pair of taxa using modified Euclidean distance (homozygote is 100% similar to itself and heterozygote is 50% similar to itself) and then estimates tree using neighbor joining algorithm (Glaubitz, et al., 2014). The tree was visualized using the web based software Interactive Tree of Life (Letunic and Bork, 2016).

Nucleotide diversity (θ_π), Tajima's D, and genetic differentiation (F_{st}) was estimated from the final SNP matrix with the *vcftools* program (Danecek, et al., 2011) using a non-overlapping sliding window of 100 kb. The window size of 100 kb was chosen to avoid sampling error that could arise from variabilities in SNP marker distribution when low coverage sequencing is used for genotyping (Gusnanto, et al., 2014). Minor allele frequency for each SNP site was also calculated using *vcftools*. Linkage disequilibrium was calculated for pairs of alleles using a sliding window of 50 SNPs in TASSEL, and decay of LD with distance was evaluated using non-linear regression *nls* function in R (R Development Core Team, 2016) with a maximum iteration of 100. Expected values of squared allele-frequency correlation (r^2) under drift equilibrium was calculated using the equation from Hill and Weir (1988) as explained in Remington, et al. (2001). Then, average r^2 and average LD half decay distance (bp) were

calculated. Nei's expected heterozygosity was calculated on a per locus basis using the *heterozygosity* function in the R package Pegas (Nei, 1987, Paradis, 2010). Average expected heterozygosity was calculated as mean of heterozygosity across all polymorphic sites.

Genomic prediction and heritability

Statistical model for prediction

Genomic best linear unbiased prediction (GBLUP) model was implemented using *kin.blup* function in R package rrBLUP (Endelman, 2011). In the GBLUP model:

$$y = \mu + g + e, \quad (2)$$

y is a vector of phenotype BLUEs; μ is the overall mean; **g** is a vector of random effect of genotypes with $g \sim N(0, \boldsymbol{A}\sigma_g^2)$, where σ_g^2 is additive genetic variance and $\boldsymbol{A}$ is the realized additive relationship matrix calculated from $n \times m$ genotype matrix with n number of genotypes and m number of markers using *A.mat* function from rrBLUP package (Endelman and Jannink, 2012); and **e** is a vector of residuals that are identical and independently distributed with $e \sim N(0, \boldsymbol{I}\sigma_e^2)$, where σ_e^2 is the residual variance and $\boldsymbol{I}$ is an identity matrix.

Estimation of genomic heritability

A re-parameterization of GBLUP model as explained in Janss, et al. (2012) was done to evaluate the impacts of population structure on genomic heritability of the traits. The reparameterized model can be written as:

$$y = 1\mu + U\alpha + e, \quad (3)$$

In the above equation, **U** is an $n \times (n-1)$ matrix of the eigenvectors obtained from eigenvalue decomposition of additive relationship matrix ($\boldsymbol{A}$) with $\mathbf{U}_i$ the column i ($i = 1, 2, \ldots, n-1$) of **U** representing the principal component loads; α is an $(n-1) \times 1$ vector of random effects with normal distribution $N(0, \mathbf{D}\sigma_g^2)$ where **D** is an $(n-1) \times (n-1)$ diagonal matrix with

each diagonal element representing eigenvalues of $\boldsymbol{A}$ corresponding to that particular column. The model (3) with principal components as random variables generates the same marker distribution as model (2), and allows for separation of total genetic variance σ_g^2 into across-subpopulation genetic variance σ_{gA}^2 due to population structure, and within-subpopulation genetic variance σ_{gW}^2. This partitioning of total genetic variance allowed for estimation of within (h_{gA}^2) and across-subpopulation (h_{gW}^2) genomic heritabilities which were calculated as:

$$h_{gA}^2 = \frac{\frac{1}{n}\sum_{i=1}^{d}\alpha_i^2}{\frac{1}{n}\sum_{i=1}^{n}\alpha_i^2 + \sigma_e^2}$$

and

$$h_{gW}^2 = \frac{\frac{1}{n}\sum_{i=d+1}^{n}\alpha_i^2}{\frac{1}{n}\sum_{i=1}^{n}\alpha_i^2 + \sigma_e^2}$$

where *d* is largest eigenvectors in the population with *n* individuals used to account for population substructure that result in artifact variation arising due to population admixture, *d* was calculated using the *eigen* function on relationship matrix, $\boldsymbol{A}$. The posterior values for σ_{gA}^2, σ_{gW}^2, σ_g^2, and σ_e^2 were estimated by Markov Chain Monte Carlo (MCMC) using a Gibbs sampler as proposed by de los Campos, et al. (2010) and Janss, et al. (2012) for each trait using phenotypic and genotypic data for all individuals in our panel. A total of 37,000 MCMC iterations were run with first 2000 iterations discarded for burn-in. The posterior means for within and across-subpopulation heritabilities were calculated form the estimated variance components.

Cross validation and prediction accuracy

Cross validation using stratified sampling (CV1)

A common cross validation approach using five-folds obtained by stratified sampling was done for CV method 1 (CV1). In stratified sampling, the individuals were proportionally sampled from each sorghum race to form cross-validation folds that have population structure similar to that of the whole population. As illustrated in Figure 1A, individuals within each race were randomly partitioned into five mutually exclusive groups (W_1, W_2, W_3, W_4, and W_5) with similar sample sizes resulting in five proportionally divided datasets (one per race). Then, five subsets (S_1, S_2, S_3, S_4, and S_5) were constructed such that each subset contained one of the partitions from each race (Figure 1A). During cross validation, each subset was treated as a fold, and four of the folds were assigned to the training set and the genetic values were predicted for the remaining fold. This process was repeated until every single fold and individuals were predicted only once, and the predicted genetic values for all individuals were stored. Prediction accuracy was calculated as correlation between predicted genetic values and observed phenotypic values of all individuals in the population for each cross validation run. A similar approach has previously been applied to study the effect of population structure in prediction results (Albrecht, et al., 2011, Guo, et al., 2014). Since all cross-validation folds are proportionally sampled from structured subpopulations, the training and validation sets used in prediction have similar racial structure.

The accuracy from CV1 method was decomposed into covariances resulting from conditional expectations due to racial structure. The decomposition of covariance was calculated as described in Sorensen and Gianola (2007):

$$\mathrm{Cov}(x, y) = \mathrm{E}_{\mathrm{race}}\,[\mathrm{Cov}(x, y|\mathrm{race})] + \mathrm{Cov}_{\mathrm{race}}\,[\mathrm{E}(x|\mathrm{race}), \mathrm{E}(y|\mathrm{race})] \quad (4)$$

where, x and y are predicted and observed values, respectively; E_{race} is expectation over races of the covariances within race; and Cov_{race} is covariance across races of the expectation within race. A multi-response model with unstructured variances was fitted with scaled values of x and y as response variables, and race as a random variable in the model using the *MCMCglmm* function in the R package MCMCglmm (Hadfield, 2010). A total of 13,000 iterations were done with 3,000 burn-ins, and posterior mean was calculated from a total of 1000 estimates were recorded using a thinning interval of 10.

Figure 1. Examples for cross-validation approaches implemented in the sorghum diversity panel. A) CV1, individuals in each race were proportionally divided into five datasets (W_1,…,W_5) and cross validation fold (S_1,…,S_5) was created as shown by rectangular boxes with broken lines ; B) CV2: within race (WR, I) and across race (AR, II) cross validation method for mixed race. A variation of AR prediction, single race training (SRT), was also implemented where a single race was used for training instead of all four races. In parentheses are the number of individuals used in prediction.

Across and within race cross validations (CV2)

While the CV1 method simulates similar population structure across both training and validation populations, breeding populations are often derived from genetically distinct pedigrees with dissimilar population structure. In order to understand how the GBLUP model for grain yield related traits in sorghum is affected by intrinsic racial structure, we designed a second cross validation experiment, CV2. In this approach, we ran predictions either by dividing individuals from a single race into training and validation folds, or by using individuals from certain race/s as a training population to predict genetic values of individuals from unrelated race/s (Fig 1B). Similar strategies have previously been reported for within and across group genomic prediction

for diversity panels in maize and rice (Guo, et al., 2014), and for breeding population in wheat (Norman, et al., 2018).

The first CV2 method, within race (WR) prediction, was done by randomly dividing individuals within a single race into five proportional folds (Fig 1B). The five-folds are used for five-fold cross validation, the predicted values for individuals in each fold was stored and subsequently a single *r* was calculated for each cross validation run, as previously described for CV1 method. The five folds in this method are derived from the five mutually exclusive datasets (W_1, W_2, W_3, W_4, and W_5) that were used in CV1 for each race; however, in this method, four of these datasets/folds collectively formed the training set and the remaining dataset/fold was used as validation set. For each run, the predicted values were stored until each fold was predicted once, then a single *r* for calculated as correlation between predicted and observed phenotypic values for the given race.

In the second CV2 method, across race (AR) prediction was conducted using four of the races as training population and the fifth race as a validation race (Fig 1B). Unlike in CV1, AR doesn't have a uniform population structure across the folds and the individuals in the training and validation populations are from genetically distinct racial clusters. In order to maintain the same training population size between AR and WR predictions, we sampled proportional amount of individuals from each race to makeup the total cross-validation sample size equal to the sum total of individuals within the validation race. For example as shown in Figure 1B, a total of 13 individuals from each of the four races kafir, caudatum, durra, and guinea were sampled as training population ($n = 13 \times 4$) and breeding values were estimated for a random sample of 13 individuals from the mixed race. Subsequently, *r* for the mixed race was calculated as correlation for observed phenotypic values and predicted genetic values for those 13 randomly sampled

individuals from the mixed race. Similarly, *r* for AR prediction of each race was calculated in similar fashion with a total of 13, 16, 27, and seven individuals sampled from each race for prediction of the races kafir, durra, caudatum, and guinea, respectively. In addition to AR method we also ran a variation of across subpopulation prediction, which we call single race training (SRT) method, where a single fold of 36 individuals sampled from a single race was used as the training population to obtain predicted genetic values of individuals from all other races (Fig 1B). The prediction accuracies for the SRT method were calculated as correlations between predicted and observed values for pairwise combination of training and validation races. The objective of this method was to explore the predictive relationship between any two races for a given trait.

A total of 100 random replications were conducted for each cross-validation method, and estimates for mean *r* and standard deviations were calculated. Vectoral graphs used in the analysis of results were created using various plotting functions in R package ggplot2 (Wickham, 2016).

RESULTS

Racial structure

We identified an optimum of five subpopulation cluster based on estimates of cross-validation error from admixture (Supplementary Figure 1). Admixture ancestry coefficients (Q) were used to assign individuals into subpopulations, individuals with coefficients >50% were assigned into that subpopulation. Four of the five subpopulation clusters, thus identified, were broadly congruent with original racial classification of the accessions based on morphological characteristics (Figure 2a, Supplementary Datafile). The remaining accessions contained mixed ancestry based on admixture components, and a large proportion of them belonged to

intermediate or mixed races based on original morphological classification (Casa, et al., 2008). For ease, the subpopulation clusters are referred to as corresponding race and "mixed" race represents the cluster of accessions with mixed or intermediate ancestry. Our results from admixture were supported by the principal component and neighbor joining analyses (Figure 2). In the neighbor joining tree, *S. propinquum*, an outgroup individual which is a diploid wild sorghum from southeast Asia, clustered together with accessions from the race guinea suggesting potentially earlier adaptation and divergence of guinea race compared to caudatum, durra, and kafir (Figure 2c).

Figure 2. Population structure and clustering analysis of the sorghum diversity panel based on; a) ancestry coefficients for K=5 in admixture, b) principal component analysis of the first three PCs, and c) neighbor joining tree analysis. In parentheses, proportion of variation explained by the corresponding PC. Branches and labels in the tree and accessions in PCA are color coded by the race identified from population structure analysis using admixture. Branch represented by broken line in the guinea clade is wild sorghum *S. propinquum*.

Genetic diversity and linkage disequilibrium

Tajima's D and heterozygosity estimates suggest presence of strong genetic bottlenecks in our panel possibly from domestication, adaptation and artificial selection. The average distance over which LD decayed to half of its maximum value was ~20 kb in our panel, which is consistent with previous observations (Hamblin, et al., 2004, Mace, et al., 2013). Whereas, the average distance for LD decay to reach background levels ($r^2 < 0.1$) was around 100 kb, similar to previous observations of Morris, et al. (2013) in global diversity panels. We observed variability in the level of genetic diversity and LD among sorghum races (Table 1). Average expected heterozygosity for all races were significantly different (p-value < 0.01) from each other. The

373 presence of extensive LD, lower genetic diversity and Tajima's D values within kafir suggests
374 the presence of a stronger genetic bottleneck within this race compared to others (Table 1).
375 Among the five racial types, the mixed race had highest average nucleotide diversity (3.5×10^{-5})
376 and lowest LD. The race guinea had the highest average expected heterozygosity (0.5), whereas
377 the average distance to LD decay for guinea were higher than all races except kafir. Genetic
378 diversity and LD for durra were similar to that of the whole panel. Although LD and nucleotide
379 diversity estimates of caudatum were comparable to that of guinea, heterozygosity in caudatum
380 was about 30% of guinea (Table 1). We calculated the Euclidean distance between the centroids
381 of five PCs and also estimated F_{st} for different races and found that kafir and mixed were
382 genetically the most and least distant race, respectively, whereas the other three races (caudatum,
383 durra, and guinea) seemed to be roughly equidistant from each other (Supplementary Table S1).
384 These results show consistency with the timeline of diversification of these races, as kafir is
385 probably the most recent and mixed race has some of the most primitive accessions of
386 intermediate and bicolor race (Deu, et al., 2006, Doggett, 1988, Kimber, et al., 2013).

387 **Table 1**. Summary statistics of whole genome estimates for genetic diversity and LD.

	Whole panel	Mixed	Kafir	Durra	Caudatum	Guinea
Number of accessions	389	67	65	82	137	38
Average heterozygosity[a]	0.17	0.19	0.12	0.18	0.14	0.5
Nucleotide diversity (10^{-5})	3.23	3.48	2.32	3.07	3.01	3.12
Tajima's D	-0.63	-0.93	-1.2	-0.88	-1.01	-0.92
Average r^2	0.09	0.1	0.21	0.1	0.11	0.16
LD decay distance (bp)	20491	14625	145252	19870	32076	39712

388 [a]Nei's unbiased estimator of gene diversity (Nei, 1987)

Phenotypic variation and correlation

The differences between the population means of at least some of the races were significantly different (p-value < 10^{-5}) for all traits except FLH and PH. Phenotypic mean and standard deviation for individual races are listed in Supplementary Table S2. Variation in phenotypic distribution and correlation between traits was observed across all traits (Figure 3). Grain yield was positively correlated with both GN and GW while the two yield components (GN and GW) were slightly negatively correlated to each other. While BL showed a significantly negative correlation with grain yield traits, remaining plant and inflorescence architecture traits (PH, FLH and PL) didn't exhibit any significant correlation to grain yield traits. Grain yield, GN, PH, FLH were all significantly positively correlated with DTA, whereas PL and GW showed significantly negative correlation with DTA.

Figure 3. Distribution and pairwise correlations for adjusted phenotypic mean for all eight traits. Histograms for traits is displayed along the diagonal. Scatterplots with line of fit (red line) for all individuals in the diversity panel are to the left and below the diagonal. Pearson correlation coefficient between the traits shown above the diagonal and to the right. Significance level: *p < 0.05, **p < 0.01, **p < 0.001. BL: panicle branch length, DTA: days to anthesis, FLH: flag leaf height, GN: grain number per primary panicle, GW: thousand grain weight, GY: grain yield per primary panicle, PH: plant height, PL: panicle length.

Prediction accuracy and racial structure

Mean prediction accuracy of a trait is known to be directly related to its heritability (Combs and Bernardo, 2013). We were interested in the nature of relationship between CV1 prediction accuracy and total genomic heritability across all traits. Therefore, we ran correlation using mean estimates of CV1 accuracy from all traits to their respective genomic heritabilities and observed

a strong correlation (0.63) between the two. For the CV1 method, the highest and lowest *r* were 0.69 for GW and 0.52 for GN and DTA, respectively. Among the traits studied, PH and GN had the highest and lowest genomic heritabilities, respectively (Figure 4a). Despite low heritability, GY had a mean *r* of 0.57 for the CV1 method and DTA had lowest *r* (0.52) despite moderate-high genomic heritability (0.73). As expected, *r* from the CV1 method were always higher than *r* from CV2 prediction methods for all traits, which can be ascribed to the larger training population size and similar population structure between training and validation population in CV1 (Table 2).

Figure 4. Posterior means of, **a**) within-subpopulation and across-subpopulation genomic heritabilities using first five principal components, **b**) scaled covariances due to condition expectation of race in CV1 prediction. *E_race* represents covariance due to race, and *Cov_race* represents covariance due to individuals within race. h^2, genomic heritability; PH, plant height; GW, thousand grain weight; FLH, flag leaf height; GN, grain number per panicle; GY, grain yield per primary panicle; DTA, days to anthesis; BL, primary branch length; PL, panicle length.

Table 2. Mean prediction accuracy (*r*) of different cross validation methods for all traits studied. Values represent mean ± standard deviation. CV1: cross validation method-1, AR: across race, WR: within race, SRT: single race training.

Trait	CV1	WR	AR	SRT
Days to anthesis (DTA)	0.52 ± 0.01	0.24 ± 0.23	0.12 ± 0.32	0.11 ± 0.23
Flag leaf height (FLH)	0.58 ± 0.03	0.41 ± 0.23	0.34 ± 0.33	0.32 ± 0.21
Grain number/panicle (GN)	0.52 ± 0.02	0.25 ± 0.12	0.27 ± 0.30	0.26 ± 0.20
1000-grain weight (GW)	0.69 ± 0.02	0.61 ± 0.10	0.37 ± 0.27	0.43 ± 0.20
Grain yield/panicle (GY)	0.57 ± 0.02	0.36 ± 0.12	0.35 ± 0.30	0.38 ± 0.17
Plant height (PH)	0.63 ± 0.02	0.46 ± 0.15	0.45 ± 0.28	0.36 ± 0.18
Panicle length (PL)	0.65 ± 0.01	0.25 ± 0.31	0.12 ± 0.33	0.14 ± 0.21
Terminal branch length (BL)	0.67 ± 0.07	0.38 ± 0.24	0.20 ± 0.31	0.26 ± 0.19

Figure 4b shows decomposition of the CV1 accuracy into covariances resulting from conditional expectation of races. The scaled covariances E_{race} and Cov_{race} represent expectation due to race

and covariances due to individuals within race, respectively. The mean covariances E_{race} and Cov_{race} were positively correlated with posterior means of across (0.61) and within race (0.81) genomic heritabilities, respectively. The estimates for covariances E_{race} and Cov_{race} were comparable to estimates of *r* for AR and WR prediction, respectively, except for height. The estimates of covariances and variance of predicted values for AR prediction method were smaller than in WR (Supplementary Table S3). Hence, the differences in mean *r* between the two methods weren't as pronounced as seen in rice and maize (Guo, et al., 2014).

Among CV2 prediction methods, estimates of *r* for WR were higher than AR for BL, DTA, FLH, GW and PL, but the two methods had similar estimates for PH, GN and GY (Table 2). However, *r* for different CV2 methods varied depending on the combination of trait and race (Supplementary Table S4). Posterior means of within-subpopulation genomic heritabilities were moderately correlated (0.4) to mean *r* from WR prediction. The average *r* for WR were higher than AR for caudatum and durra whereas the two methods had similar averages for guinea, kafir, and mixed (Supplementary Table S4). Plant height and FLH showed smaller difference in *r* between AR and WR prediction across all races, whereas GW showed consistently higher estimates of *r* for WR over AR for all races. Among all the traits, GW had highest mean *r* for WR (0.61) and SRT (0.43) prediction methods, while PH (0.45) had highest *r* for AR.

In SRT method, the traits that are heavily correlated to racial structure (DTA, PL, BL, GN and GY) were poorly predicted than PH, FLH and GW (Figure 5). In general, the races durra and caudatum resulted in poor prediction for SRT method when introduced in model as training or validation populations compared to mixed and guinea races (Figure 5). The two former races are thought to have diverged recently than the latter two (Kimber, et al., 2013). So we conducted an additional across race cross-validation using a random subset of 36 kafir accessions as

validation population and a combination of 36 accessions from one or many remaining races as training population (Supplementary Figure S2). We started with 36 accessions from mixed race based on earlier divergence and best predictor of kafir in CV2 SRT prediction results. Mixed race by itself predicted as good as or better than most of combination which is expected due to the closer relationship and larger diversity of the race (Supplementary Figure S2). However, a combination of guinea and mixed performed the best for plant height and grain number. Grain yield showed increase in accuracy with combination of various races except in the case of MD (mixed-durra). While mixed race is more closely related to all other races, consistently better performance of guinea and kafir as validation population in SRT could be due the amount of shared (versus population-specific) allelic variation in these groups. We observed higher proportions of intermediate frequency minor alleles (0.1 to 0.4) in mixed and guinea than in caudatum, durra and kafir (Supplementary Figure S3). We identified private alleles in each races as the minor alleles that were present in a particular race and were absent in all other races. Within the mixed race group, there were 3,378 private polymorphisms that were not present in any of the other races, although on average these were only present at low frequencies within the population mean (MAF = 0.03). Caudatum had fewer private polymorphisms (1,843), with a mean MAF of 0.04. Durra, on the other hand, had the highest number of private SNPs with 3,969 and the highest mean MAF for these sites (0.07). The races kafir and guinea had no private alleles.

Figure 5. Heatmap showing mean prediction accuracies (*r*) from pairwise single race (SRT) prediction in CV2 prediction method. The races to the right of the heatmap represent training race followed by validation race. Tree cluster to the top and left is based on hierarchical clustering of the values from column and rows, respectively. PH, plant height; GW, thousand

grain weight; FLH, flag leaf height; GN, grain number per panicle; GY, grain yield per primary panicle; DTA, days to anthesis; BL, primary branch length; PL, panicle length.

In order to assess the effect of training population size in *r* for AR and WR method, we ran cross validations using accessions in race caudatum using training population sizes of 28, 52, 74, 96, and 110. We ran cross validations only for PH, GN, GW, and GY because they have varying trait genetic architecture, PH and GW have relatively high genomic heritability, whereas GY and GN have relatively low heritability. So we reasoned using these four traits will be sufficient to deduce necessary information about the role of training population size, while keeping the analysis relatively simple. We observed that while increasing training population size showed consistent increase in *r* for WR prediction of all four traits, increasing training population size did not always lead to increased *r* in AR prediction (Figure 6). Mean *r* for WR was always higher than AR for GW across all training population sizes, whereas they were similar for PH among the two methods. For GN and GY, *r* was higher for AR prediction when training population size was 28 individuals and similar when training population size was 52. At larger training population size, WR prediction method had larger *r* than AR prediction.

Figure 6. Mean prediction accuracies from across race (AR) and within race (WR) prediction methods for different training population sizes in caudatum. Colors represent cross validation methods, blue = AR, red = WR. GN = grain number, GW = grain weight, GY = grain yield, PH = plant height.

DISCUSSION

Genetic differentiation and racial structure

The results from the population structure analysis support five subpopulation clusters in our sorghum diversity panel. While the four subpopulations were congruent with the four most

recent sorghum races, the bicolor race was not diverged enough to form one distinct group. This is consistent with previous observations from clustering analysis in populations consisting of diverse sorghum accessions (Brown, et al., 2011, Deu, et al., 2006, Wang, et al., 2013). The hypothesis of Harlan and Stemler (1976) that the guinea race was probably the earliest to have diverged from the early bicolor domesticates is supported by our neighbor joining analysis (Figure 2c). *Sorghum propinquum*, an outgroup diploid species from southeast Asia, clustering together with the guinea clade suggests guinea potentially diverged from the early bicolor prior to the divergence of kafir, durra and caudatum. Furthermore, we also see a large proportion of shared alleles and higher allelic diversity among guinea than the evolutionarily recent races, which, however, could also be due to higher rate of gene flow.

Genetic diversity and linkage disequilibrium

Our diversity panel is representative of much of the genetic, phenotypic and geographic diversity of the global sorghum germplasm, and therefore it is an excellent resource for genetic dissection of agronomically important traits and adaptation. Patterns of LD, Tajima's D and genetic diversity from the whole population and within different races suggest strong genetic bottlenecks in our population as a result of domestication, adaptation, and artificial selection. The rate of LD decay, Tajima's D and genetic diversity in our population were comparable to estimates from previous studies (Hamblin, et al., 2004, Mace, et al., 2013, Wang, et al., 2013). Average distance to half decay of LD was similar to that reported by Mace, et al. (2013) for improved inbreds, and lower than the previous estimates of Morris, et al. (2013). Wang, et al. (2013) using 242 diverse accessions from a sorghum mini core collection and 13,390 SNPs, found that LD decayed to background levels between 10 and 30 kb for all but chromosome 2. Average r^2 for different sorghum races in our study were consistent with estimates of Wang et al. (2013). Genetically

523 least diverse racial types (kafir and caudatum) showed higher extents of LD than the races with
524 higher diversity, with the exception of guinea. Previous studies involving separate diversity
525 panels have also reported reduced diversity in kafir (Deu et al. 2006; Casa et al. 2008). The
526 extensive LD and lower heterozygosity in kafir and caudatum might be the result of limited
527 cross-pollination and geographical isolation, whereas genetic drift and smaller sample size could
528 have contributed to slower rate of LD decay in guinea. Bouchet, et al. (2012) also observed
529 similar outcomes for LD and genetic diversity among sorghum races. The lower abundance of
530 private alleles among races kafir and guinea than caudatum and durra was observed by Bouchet,
531 et al. (2012), which is consistent with lack of private alleles for guinea and kafir in our
532 population. Despite being the oldest and most heterozygous race guinea didn't possess any
533 private allele, which could be a result of small sample size of guinea in our population.

534 **Genomic prediction and racial structure**

535 Heritabilities and CV1 prediction accuracy for grain yield and plant height in our study were
536 similar to previous studies in sorghum (Fernandes, et al., 2018, Hunt, et al., 2018, Yu, et al.,
537 2016). Similar heritabilities and *r* for flowering time, panicle length, plant height and grain
538 number have previously been reported in a rice diversity panel (Guo, et al., 2014) . The
539 overwhelming contribution of within-subpopulation genomic heritabilities towards the total
540 genomic heritabilities of traits is comparable to previous observation in rice (Guo, et al., 2014).
541 Grain yield was predicted with consistently higher accuracy than grain number across all cross
542 validation methods despite similar heritabilities of the two traits. While both GN and GY are
543 highly correlated complex traits controlled by a large number of small effect loci, higher
544 accuracy of GY over GN could have resulted from strong positive correlation (0.43) of GY to
545 GW, which has the highest *r*. Traits controlled by large number of small effect loci are predicted

546 with higher accuracy when higher allelic diversity exists in the training population as compared
547 to traits governed by few relatively large effect loci (Norman, et al., 2018). With smaller training
548 sizes for GN and GY the breadth of genetic diversity from all races might have led to boost in *r*
549 for AR over WR, but as training size increased the genomic relatedness in WR appeared to
550 outweigh the effect of genetic diversity, which resulted in stronger positive relationship between
551 *r* and training size in WR (Figure 6). Since the range of training size in our study is small,
552 genetic diversity did not increase substantially with increasing training size resulting in lack of
553 linear relationship between training size and *r* in AR for GN and GY.

554 Population structure resulting from domestication and diversifying selection leads to
555 varying levels of genetic relatedness among individuals within and between subpopulations. The
556 accuracy with which breeding values are estimated is affected by stratification in the population,
557 and the effects are more pronounced when the genetic architecture of the predicted trait is
558 directly associated with population structure (Isidro, et al., 2015, Windhausen, et al., 2012).
559 Previous studies have shown that when population structure in training and testing populations is
560 similar, it can contribute positively towards prediction accuracy (Bastiaansen, et al., 2012,
561 Crossa, et al., 2014, Guo, et al., 2014). But when cross validation strategies that constrained
562 population structure were implemented, it resulted in decline of *r* due to the weakened genetic
563 relationship among individuals in training and testing population (Guo, et al., 2014, Ly, et al.,
564 2013, Lyra, et al., 2018, Norman, et al., 2018).

565 In order to understand the contribution of racial structure in prediction accuracy of
566 stratified sampling method, we decomposed the CV1 accuracy into expectation over races (E_{race})
567 and covariance due to individuals within race (Cov_{race}). Almost non-existent E_{race} covariances for
568 the two height traits, FLH and PH, indicates race as a predictor contributed poorly towards

prediction of height. On the other hand, race contributed relatively larger proportion of total covariance for grain yield components and panicle architecture traits. Since the racial structure of sorghum can be directly associated with panicle architecture and indirectly to the grain yield components, proportion of across race genomic heritability and covariance due to race were higher for those traits. We saw sharper decline in *r* for AR prediction compared to WR, especially for panicle architecture traits, which could be attributed to poor genomic relationship between training and validation population. Height traits, PH and FLH, that are less associated with racial structure showed smaller decline in *r* than panicle architecture and grain yield traits. The SRT prediction method also showed smaller *r* for pairwise prediction results for the panicle architecture traits than other traits. Yu, et al. (2016) have previously observed that race as a predictor explains higher variation in predicted values of biomass traits than actual phenotypic values in sorghum, suggesting that under the presence of similar racial structure in training and validation population the accuracy of genomic prediction might have been inflated as a result of overemphasis on racial differences (Brown, 2016). Our approach of decomposition of covariances into conditional expectations due to race could be utilized in dissection of impact of population structure in cross validation accuracy from stratified and random sampling methods in diverse as well as breeding populations.

Cross validation approaches similar to the one employed in this study have resulted in higher *r* for within population prediction than across population prediction in wheat (Norman, et al., 2018), rice and maize (Guo, et al., 2014). Although average *r* across all races in our study was higher for within population for most of the traits, the variation in *r* for individual race and trait combination shows interaction between population structure and trait genetic architecture. Higher *r* for WR over AR among the races caudatum and durra could be because of higher

proportion of private alleles and smaller proportion of intermediate frequency minor alleles. This could be the reason for smaller difference between average *r* for AR and WR among guinea, kafir and mixed, as these races show lack of private alleles and/or higher genetic diversity. Furthermore, the results from clustering analysis have shown that the mixed race has closest genetic relationship to the rest of the four races. Our SRT prediction which was a good measure of pairwise predictive relationship between two races also shows that kafir, guinea and mixed have better predictive relationship to each other than to caudatum and durra. This was further supported by our cross-validation using various combination of races to predict kafir, evolutionarily the most recent race (Supplementary Figure S2). Genetic diversity and divergence seems to have an important impact in prediction accuracy, which needs to be an important consideration during training population design for diverse germplasm evaluation.

Potential applications for sorghum breeding

Genomic prediction was first introduced roughly two decades ago and has been applied in plant breeding for over a decade (Bernardo and Yu, 2007, Meuwissen, et al., 2001). However, studies investigating prospects and applications of genomic prediction in sorghum are limited. A few studies have been reported for biomass traits in diversity panels (Yu et al. 2016; Fernandes et al. 2017), grain yield in pedigreed male inbred lines (Hunt et al. 2018), and a simulation study investigating prospects in a small sorghum breeding program (Muleta et al. 2019). While clearly defined heterotic pools do not exist in sorghum as they do in maize, races have long been exploited by sorghum breeding programs for hybrid production. If we are to exploit the vast genetic diversity of races in sorghum breeding, we need a more comprehensive understanding of how racial structure impacts prediction accuracy for economically and agronomically important traits.

Our study suggests maintaining a genetically diverse training population that includes a mixed/intermediate race might boost prediction accuracy when training population size is constrained. This strategy might be beneficial for young and small breeding programs where breeders have limited resource to construct individual training population for different breeding populations (Muleta, et al., 2019). Furthermore, new phenotypic data from diverse lines when added into the training population can allow for maintenance and increase in frequency of advantageous minor alleles in the gene pool. Guinea had the highest mean prediction accuracy for grain yield and grain weight irrespective of prediction method, suggesting a genetically diverse training population is likely to predict the yield potential of best performing guinea more accurately than individuals from any other race. So breeding programs in West Africa, where guinea sorghum is widely grown, could utilize genotypic and phenotypic data from all racial types in training population design for genomic prediction of guinea varieties. Our results from the SRT method showed that moderate prediction accuracy can be gained even by using a single completely unrelated race as a training population for grain yield components and height. Historically, sorghum breeding programs in the US have heavily relied on kafir and caudatum types while genetic diversity from other races are underutilized. While breeding programs with plentiful resources could gain higher selection accuracy by simply increasing training population size and designing several independent training populations, the utilization of interracial diversity in genomic prediction could help in introducing novel sources of variation for diseases and pest resistance as well as genetic variation for increasing yield potential in the long run. Similarly, another way to increase selection accuracy and genetic gain of complex traits is through utilization of trait-assisted and indirect genomic selection when highly heritable and correlated secondary traits are available (Fernandes et al. 2017). For example, durra accessions in

our results show a within race prediction accuracy of 0.33 and 0.43 for grain number and grain yield but an accuracy of 0.81 and 0.79 for branch length and panicle length, respectively. Panicle architecture could be used in indirect or trait-assisted genomic prediction for grain yield by breeding programs dominated by durra type sorghum varieties. In addition to utilization of within and across group genetic variances, optimization algorithms could also help in efficient design of training population for diversity panels and breeding populations (Akdemir, et al., 2015, Isidro, et al., 2015).

For effective use of crop diversity in sorghum breeding, a breeder might want to work with best representatives from all races rather than opting for the best lines of some races (Brown, 2016). In practical applications, prediction accuracy of traits that are affected by population structure can be increased by using genetically distant subpopulations as parental lines (Guo, et al., 2014). Our results can be useful in such an effort because understanding how individuals of certain race respond to models trained using unrelated races can provide insights into how overall genetic diversity can be deployed in prediction of different racial types. For example, the prediction results from our SRT method shows guinea or mixed race with 37 individuals predicted GN and GY in kafir with higher accuracy than from using 52 kafir accessions (Figure 3, Supplementary Table S3). This kind of empirical evidences of predictive relationship can help in identifying trait-specific and race-specific training design for genomic prediction highlighting the need for more explorative and empirical case studies in natural and breeding populations.

CONCLUSION

Similar population structure between training and testing population can have positive impact on accuracy of genomic prediction. However, inflation in prediction accuracy could be an outcome of genomic prediction models overemphasizing racial differences. Prediction accuracy among races with higher proportion of allelic diversity and/or shared alleles is boosted by training population with higher genetic diversity despite poor genomic relationship, whereas genomic relationship outweighed genetic diversity among races with limited diversity and/or presence of unique polymorphisms. Therefore, training population design for a historically diverse and structured population in sorghum requires careful consideration of genetic structure of the testing population. While the sorghum association panel (SAP) was not intended for genomic prediction, the breadth of genetic and phenotypic diversity in this panel can allow for its application as training population for estimation of breeding values of diverse gene bank accessions. To that objective, including more guinea and bicolor accessions to the panel would be beneficial because our results show that accessions in these races boosts prediction accuracy as training and testing population.

ACKNOWLEDGEMENTS

We would like to thank the endowment fund for the Robert and Lois Coker Trustees Chair of Genetics, Wade Stackhouse fellowship, and Clemson University's Public Service and Agriculture agency for their support for our research and training. We are grateful to Jianming Yu, Jim Holland, Jean-Luc Jannink and our reviewers for their helpful comments and suggestions.

SUPPLEMENTAL MATERIAL

Supplementary_File1 consists of three supplementary figures and four supplementary tables.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Akdemir, D., J.I. Sanchez and J.-L. Jannink. 2015. Optimization of genomic selection training populations with a genetic algorithm. Genet Sel Evol 47: 38.

Albrecht, T., V. Wimmer, H.J. Auinger, M. Erbe, C. Knaak, M. Ouzunova, H. Simianer and C.C. Schon. 2011. Genome-based prediction of testcross values in maize. Theor Appl Genet 123: 339-350. doi:10.1007/s00122-011-1587-7.

Alexander, D.H., J. Novembre and K. Lange. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19: 1655-1664. doi:10.1101/gr.094052.109.

Bastiaansen, J.W., A. Coster, M.P. Calus, J.A. van Arendonk and H. Bovenhuis. 2012. Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures. Genet Sel Evol 44: 3. doi:10.1186/1297-9686-44-3.

Bates, D., M. Mächler, B. Bolker and S. Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software 67: 48. doi:10.18637/jss.v067.i01.

Bernardo, R. and J. Yu. 2007. Prospects for genomewide selection for quantitative traits in maize. Crop Science 47: 1082-1090.

Bouchet, S., D. Pot, M. Deu, J.F. Rami, C. Billot, X. Perrier, R. Rivallan, L. Gardes, L. Xia, P. Wenzl, A. Kilian and J.C. Glaszmann. 2012. Genetic structure, linkage disequilibrium and signature of selection in Sorghum: lessons from physically anchored DArT markers. PLoS One 7: e33470. doi:10.1371/journal.pone.0033470.

Boyles, R.E., E.A. Cooper, M.T. Myers, Z. Brenton, B.L. Rauh, G.P. Morris and S. Kresovich. 2016. Genome-Wide Association Studies of Grain Yield Components in Diverse Sorghum Germplasm. Plant Genome 9. doi:10.3835/plantgenome2015.09.0091.

Boyles, R.E., B.K. Pfeiffer, E.A. Cooper, B.L. Rauh, K.J. Zielinski, M.T. Myers, Z. Brenton, W.L. Rooney and S. Kresovich. 2017. Genetic dissection of sorghum grain quality traits using diverse and segregating populations. Theor Appl Genet 130: 697-716. doi:10.1007/s00122-016-2844-6.

707 Brown, P.J. 2016. Plant breeding: Effective use of genetic diversity. Nat Plants 2: 16154.
708 doi:10.1038/nplants.2016.154.

709 Brown, P.J., P.E. Klein, E. Bortiri, C.B. Acharya, W.L. Rooney and S. Kresovich. 2006. Inheritance of inflorescence
710 architecture in sorghum. Theor Appl Genet 113: 931-942. doi:10.1007/s00122-006-0352-9.

711 Brown, P.J., S. Myles and S. Kresovich. 2011. Genetic support for phenotype-based racial classification in sorghum.
712 Crop Science 51: 224-230.

713 Casa, A.M., G. Pressoir, P.J. Brown, S.E. Mitchell, W.L. Rooney, M.R. Tuinstra, C.D. Franks and S. Kresovich.
714 2008. Community resources and strategies for association mapping in sorghum. Crop Science 48: 30-40.

715 Combs, E. and R. Bernardo. 2013. Accuracy of genomewide selection for different traits with constant population
716 size, heritability, and number of markers. Plant Genome 6.

717 Crossa, J., P. Perez, J. Hickey, J. Burgueno, L. Ornella, J. Ceron-Rojas, X. Zhang, S. Dreisigacker, R. Babu, Y. Li,
718 D. Bonnett and K. Mathews. 2014. Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity
719 112: 48-60. doi:10.1038/hdy.2013.16.

720 Danecek, P., A. Auton, G. Abecasis, C.A. Albers, E. Banks, M.A. DePristo, R.E. Handsaker, G. Lunter, G.T. Marth,
721 S.T. Sherry, G. McVean and R. Durbin. 2011. The variant call format and VCFtools. Bioinformatics 27: 2156-2158.
722 doi:10.1093/bioinformatics/btr330.

723 de los Campos, G., D. Gianola, G.J. Rosa, K.A. Weigel and J. Crossa. 2010. Semi-parametric genomic-enabled
724 prediction of genetic values using reproducing kernel Hilbert spaces methods. Genetics Research 92: 295-308.

725 de los Campos, G. and D. Sorensen. 2014. On the genomic analysis of data from structured populations. J Anim
726 Breed Genet 131: 163-164. doi:10.1111/jbg.12091.

727 Deu, M., F. Rattunde and J. Chantereau. 2006. A global view of genetic diversity in cultivated sorghums using a
728 core collection. Genome 49: 168-180. doi:10.1139/g05-092.

729 Doggett, H. 1988. Sorghum. Longman Scientific and Technical, New York, N.Y.

730 Endelman, J.B. 2011. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. Plant
731 Genome 4: 250-255. doi:10.3835/plantgenome2011.08.0024.

732 Endelman, J.B. and J.L. Jannink. 2012. Shrinkage estimation of the realized relationship matrix. G3 2: 1405-1413.
733 doi:10.1534/g3.112.004259.

734 Fernandes, S.B., K.O.G. Dias, D.F. Ferreira and P.J. Brown. 2018. Efficiency of multi-trait, indirect, and trait-
735 assisted genomic selection for improvement of biomass sorghum. Theor Appl Genet 131: 747-755.
736 doi:10.1007/s00122-017-3033-y.

737 Glaubitz, J.C., T.M. Casstevens, F. Lu, J. Harriman, R.J. Elshire, Q. Sun and E.S. Buckler. 2014. TASSEL-GBS: a
738 high capacity genotyping by sequencing analysis pipeline. PLoS One 9: e90346. doi:10.1371/journal.pone.0090346.

739 Guo, Z., D.M. Tucker, C.J. Basten, H. Gandhi, E. Ersoz, B. Guo, Z. Xu, D. Wang and G. Gay. 2014. The impact of
740 population structure on genomic prediction in stratified populations. Theor Appl Genet 127: 749-762.
741 doi:10.1007/s00122-013-2255-x.

742 Gusnanto, A., C.C. Taylor, I. Nafisah, H.M. Wood, P. Rabbitts and S. Berri. 2014. Estimating optimal window size
743 for analysis of low-coverage next-generation sequence data. Bioinformatics 30: 1823-1829.
744 doi:10.1093/bioinformatics/btu123.

745 Hadfield, J.D. 2010. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R
746 package. Journal of Statistical Software 33: 1-22.

747 Hamblin, M.T., S.E. Mitchell, G.M. White, J. Gallego, R. Kukatla, R.A. Wing, A.H. Paterson and S. Kresovich.
748 2004. Comparative population genetics of the panicoid grasses: sequence polymorphism, linkage disequilibrium and
749 selection in a diverse sample of sorghum bicolor. Genetics 167: 471-483.

750 Harlan, J.R. and J.M.J. de Wet. 1972. A Simplified Classification of Cultivated Sorghum1. Crop Science 12: 172-
751 176. doi:10.2135/cropsci1972.0011183X001200020005x.

752 Harlan, J.R. and A. Stemler. 1976. The races of sorghum in Africa. In: J. R. Harlan, J. M. J. de Wet and A. Stemler,
753 editors, Origins of African plant domestication. Mouton Publishers, Paris. p. 465-478.

754 Hill, W.G. and B.S. Weir. 1988. Variances and covariances of squared linkage disequilibria in finite populations.
755 Theor Popul Biol 33: 54-78.

756 Hunt, C.H., F.A. van Eeuwijk, E.S. Mace, B.J. Hayes and D.R. Jordan. 2018. Development of Genomic Prediction
757 in Sorghum. Crop Science 58: 690-700. doi:10.2135/cropsci2017.08.0469.

758 Isidro, J., J.L. Jannink, D. Akdemir, J. Poland, N. Heslot and M.E. Sorrells. 2015. Training set optimization under
759 population structure in genomic selection. Theor Appl Genet 128: 145-158. doi:10.1007/s00122-014-2418-4.

760 Janss, L., G. de Los Campos, N. Sheehan and D. Sorensen. 2012. Inferences from genomic models in stratified
761 populations. Genetics 192: 693-704. doi:10.1534/genetics.112.141143.

762 Kimber, C.T., J.A. Dahlberg and S. Kresovich. 2013. The gene pool of Sorghum bicolor and its improvement.
763 Genomics of the Saccharinae. Springer. p. 23-41.

764 Letunic, I. and P. Bork. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of
765 phylogenetic and other trees. Nucleic Acids Res 44: W242-245. doi:10.1093/nar/gkw290.

766 Ly, D., M. Hamblin, I. Rabbi, G. Melaku, M. Bakare, H.G. Gauch, R. Okechukwu, A.G. Dixon, P. Kulakow and J.-
767 L. Jannink. 2013. Relatedness and genotype × environment interaction affect prediction accuracies in genomic
768 selection: a study in cassava. Crop Science 53: 1312-1325.

769 Lyra, D.H., Í.S.C. Granato, P.P.P. Morais, F.C. Alves, A.R.M. dos Santos, X. Yu, T. Guo, J. Yu and R. Fritsche-
770 Neto. 2018. Controlling population structure in the genomic prediction of tropical maize hybrids. Molecular
771 Breeding 38: 126. doi:10.1007/s11032-018-0882-2.

772 Mace, E.S., S. Tai, E.K. Gilding, Y. Li, P.J. Prentis, L. Bian, B.C. Campbell, W. Hu, D.J. Innes, X. Han, A.
773 Cruickshank, C. Dai, C. Frere, H. Zhang, C.H. Hunt, X. Wang, T. Shatte, M. Wang, Z. Su, J. Li, X. Lin, I.D.

774 Godwin, D.R. Jordan and J. Wang. 2013. Whole-genome sequencing reveals untapped genetic potential in Africa's
775 indigenous cereal crop sorghum. Nat Commun 4: 2320. doi:10.1038/ncomms3320.

776 McCouch, S., G.J. Baute, J. Bradeen, P. Bramel, P.K. Bretting, E. Buckler, J.M. Burke, D. Charest, S. Cloutier, G.
777 Cole, H. Dempewolf, M. Dingkuhn, C. Feuillet, P. Gepts, D. Grattapaglia, L. Guarino, S. Jackson, S. Knapp, P.
778 Langridge, A. Lawton-Rauh, Q. Lijua, C. Lusty, T. Michael, S. Myles, K. Naito, R.L. Nelson, R. Pontarollo, C.M.
779 Richards, L. Rieseberg, J. Ross-Ibarra, S. Rounsley, R.S. Hamilton, U. Schurr, N. Stein, N. Tomooka, E. van der
780 Knaap, D. van Tassel, J. Toll, J. Valls, R.K. Varshney, J. Ward, R. Waugh, P. Wenzl and D. Zamir. 2013.
781 Agriculture: Feeding the future. Nature 499: 23-24. doi:10.1038/499023a.

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

784 Morris, G.P., P. Ramu, S.P. Deshpande, C.T. Hash, T. Shah, H.D. Upadhyaya, O. Riera-Lizarazu, P.J. Brown, C.B.
785 Acharya, S.E. Mitchell, J. Harriman, J.C. Glaubitz, E.S. Buckler and S. Kresovich. 2013. Population genomic and
786 genome-wide association studies of agroclimatic traits in sorghum. Proc Natl Acad Sci 110: 453-458.
787 doi:10.1073/pnas.1215985110.

788 Muleta, K.T., G. Pressoir and G.P. Morris. 2019. Optimizing Genomic Selection for a Sorghum Breeding Program
789 in Haiti: A Simulation Study. G3 9: 391-401. doi:10.1534/g3.118.200932.

790 Nei, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York.

791 Norman, A., J. Taylor, J. Edwards and H. Kuchel. 2018. Optimising Genomic Selection in Wheat: Effect of Marker
792 Density, Population Size and Population Structure on Prediction Accuracy. G3 8: 2889-2899.
793 doi:10.1534/g3.118.200311.

794 Paradis, E. 2010. pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics
795 26: 419-420. doi:10.1093/bioinformatics/btp696.

796 Price, A.L., N.A. Zaitlen, D. Reich and N. Patterson. 2010. New approaches to population stratification in genome-
797 wide association studies. Nat Rev Genet 11: 459-463. doi:10.1038/nrg2813.

798 R Development Core Team. 2016. R: A language and environment for statistical computing. R Foundation for
799 Statistical Computing, Vienna, Austria.

800 Remington, D.L., J.M. Thornsberry, Y. Matsuoka, L.M. Wilson, S.R. Whitt, J. Doebley, S. Kresovich, M.M.
801 Goodman and E.S.t. Buckler. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize
802 genome. Proc Natl Acad Sci 98: 11479-11484. doi:10.1073/pnas.201394398.

803 Revelle, W.R. 2011. psych: Procedures for Psychological, Psychometric, and Personality Research. R package
804 version 1.1-2. Evanston, Illinois.

805 Sorensen, D. and D. Gianola. 2007. Likelihood, Bayesian, and MCMC methods in quantitative genetics.Springer.
806 (Page 67).

807 Swarts, K., H. Li, J.R. Navarro, D. An, M. Romay, S. Hearne, C. Acharya, J. Glaubitz, S. Mitchell and R. Elshire.
808 2014. FSFHap (Full-Sib Family Haplotype Imputation) and FILLIN (Fast, Inbred Line Library ImputatioN)
809 optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. Plant Genome 7: 1-
810 12.

811 Technow, F., C. Riedelsheimer, T.A. Schrag and A.E. Melchinger. 2012. Genomic prediction of hybrid performance
812 in maize with models incorporating dominance and population specific marker effects. Theor Appl Genet 125:
813 1181-1194. doi:10.1007/s00122-012-1905-8.

814 Wang, Y.H., H.D. Upadhyaya, A.M. Burrell, S.M. Sahraeian, R.R. Klein and P.E. Klein. 2013. Genetic structure
815 and linkage disequilibrium in a diverse, representative collection of the C4 model plant, Sorghum bicolor. G3 3:
816 783-793. doi:10.1534/g3.112.004861.

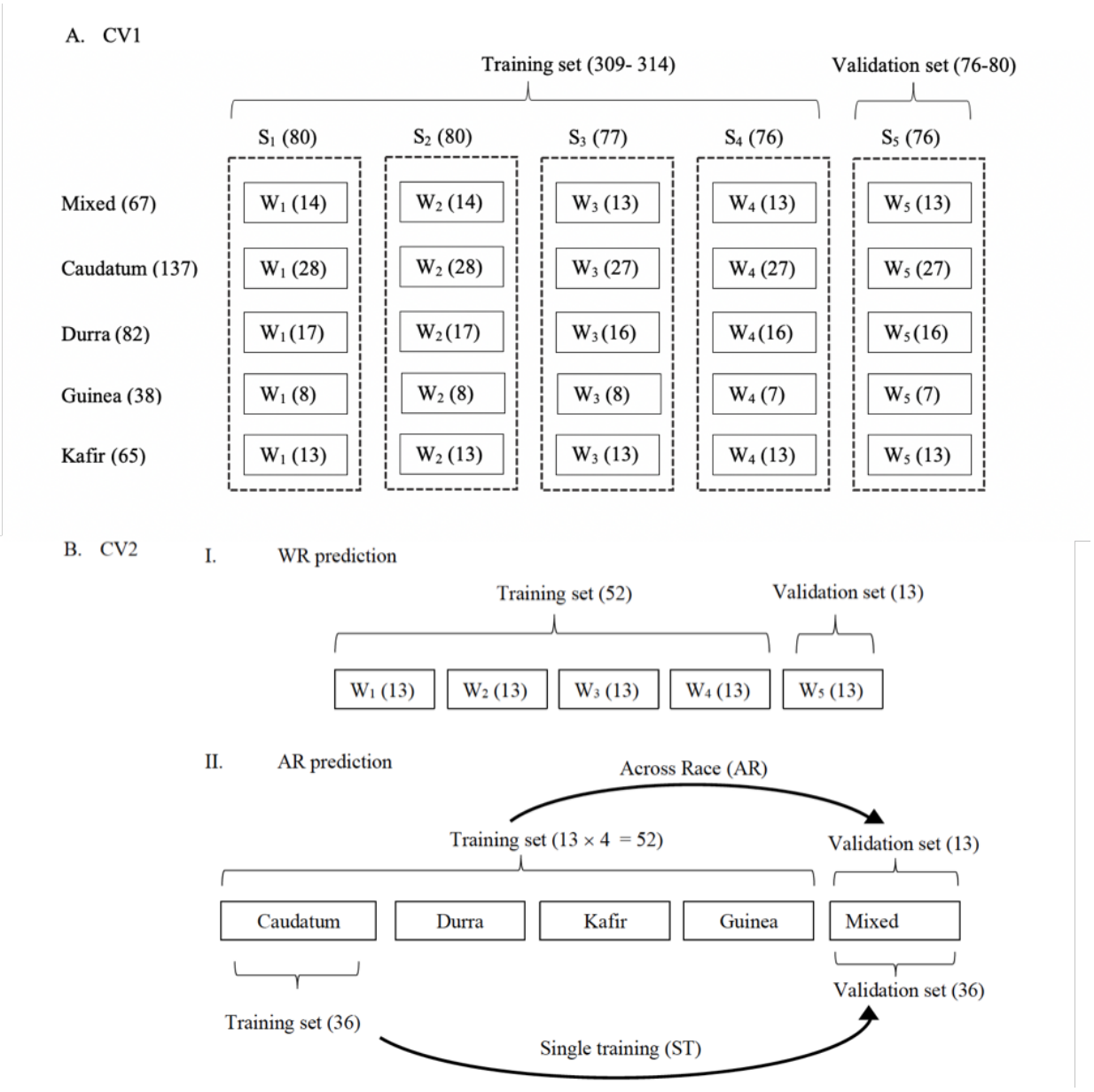
817 Wendorf, F., A.E. Close, R. Schild, K. Wasylikowa, R.A. Housley, J.R. Harlan and H. Królik. 1992. Saharan
818 exploitation of plants 8,000 years BP. Nature 359: 721-724. doi:10.1038/359721a0.

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

820 Windhausen, V.S., G.N. Atlin, J.M. Hickey, J. Crossa, J.L. Jannink, M.E. Sorrells, B. Raman, J.E. Cairns, A.
821 Tarekegne, K. Semagn, Y. Beyene, P. Grudloyma, F. Technow, C. Riedelsheimer and A.E. Melchinger. 2012.
822 Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and
823 environments. G3 2: 1427-1436. doi:10.1534/g3.112.003699.

824 Yu, X., X. Li, T. Guo, C. Zhu, Y. Wu, S.E. Mitchell, K.L. Roozeboom, D. Wang, M.L. Wang, G.A. Pederson, T.T.
825 Tesso, P.S. Schnable, R. Bernardo and J. Yu. 2016. Genomic prediction contributing to a promising global strategy
826 to turbocharge gene banks. Nat Plants 2: 16150. doi:10.1038/nplants.2016.150.
827

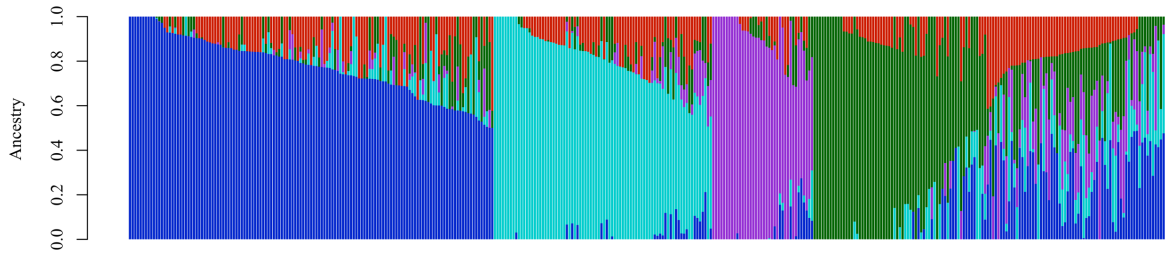
828 Figure 1.



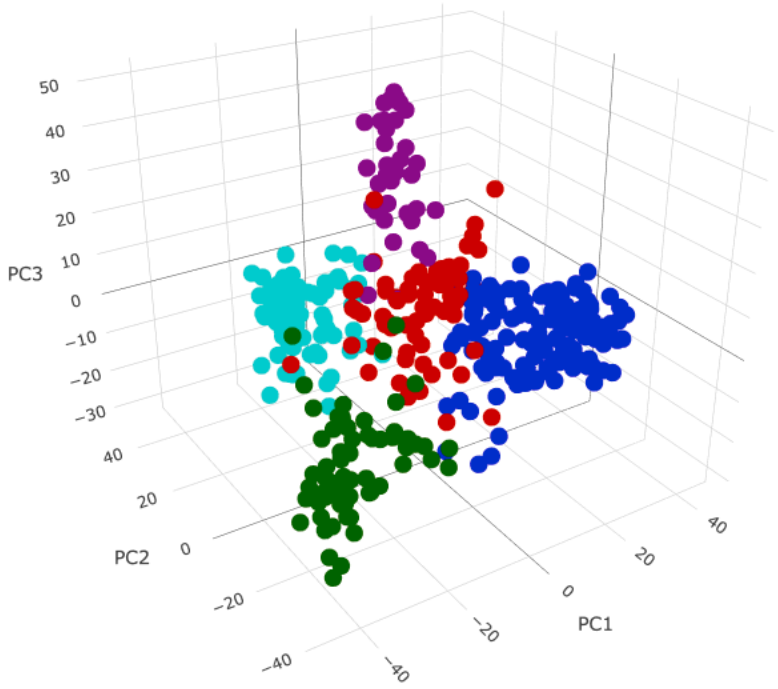
829

Figure 2.

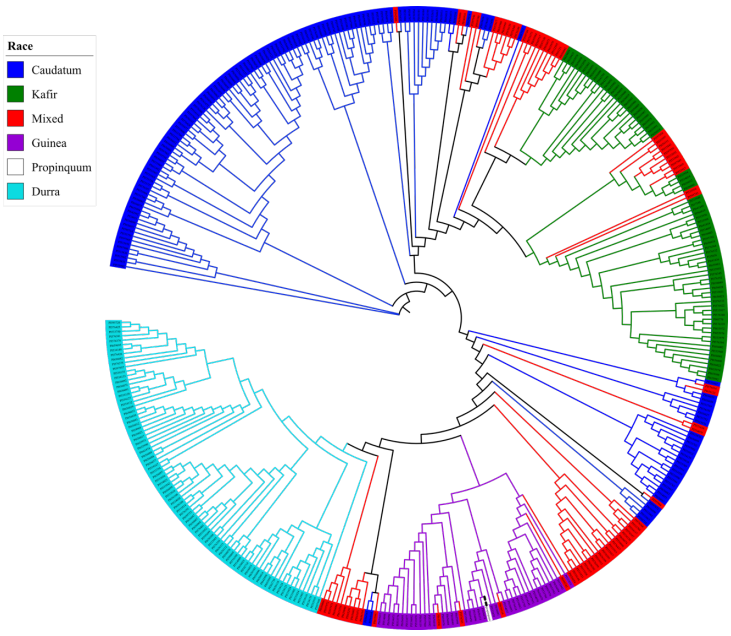
a.


Ancestry
0.0
0.2
0.4
0.6
0.8
1.0

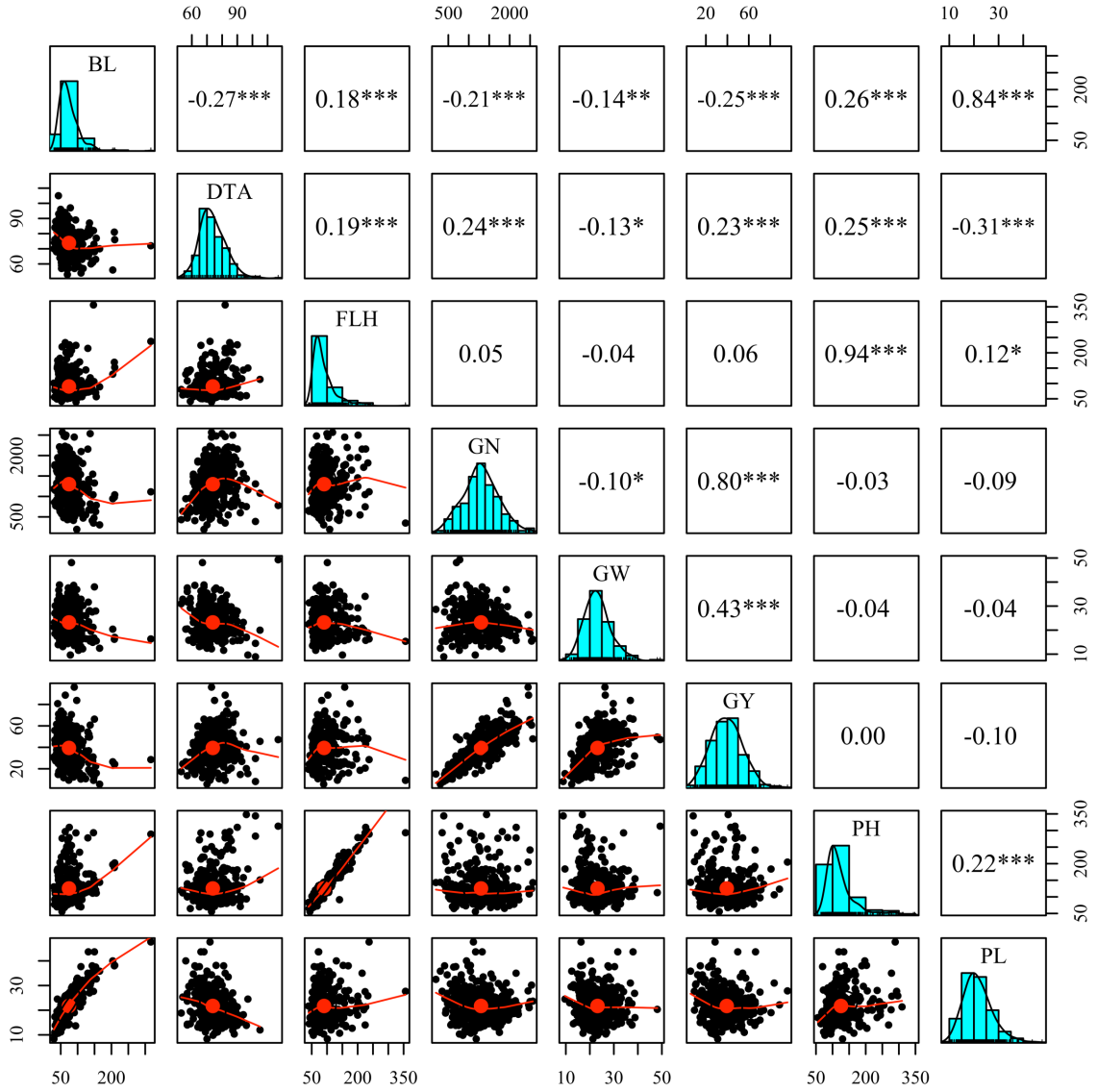
b.


PC3
50
40
30
20
10
0
−10
−20
−30
PC2
40
20
0
−20
−40
PC1
40
20
0
−20
−40

c.



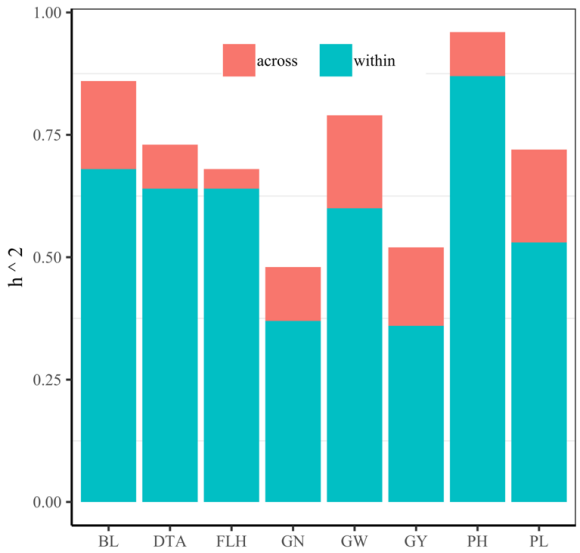
837 Figure 3.



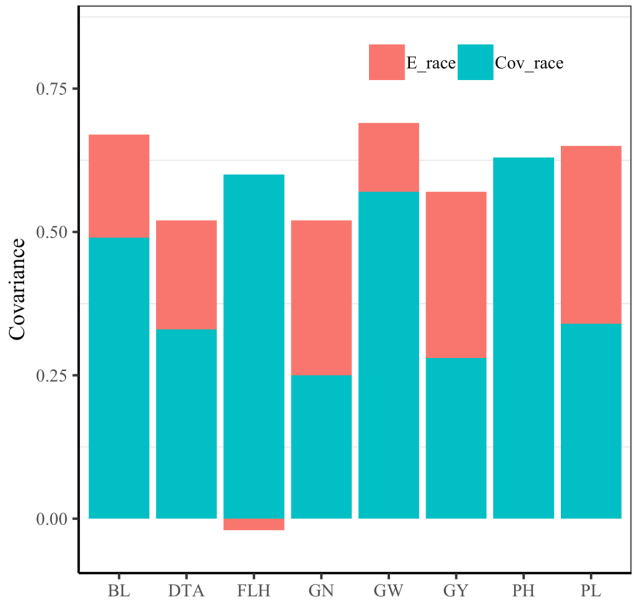
838

Figure 4.

a.

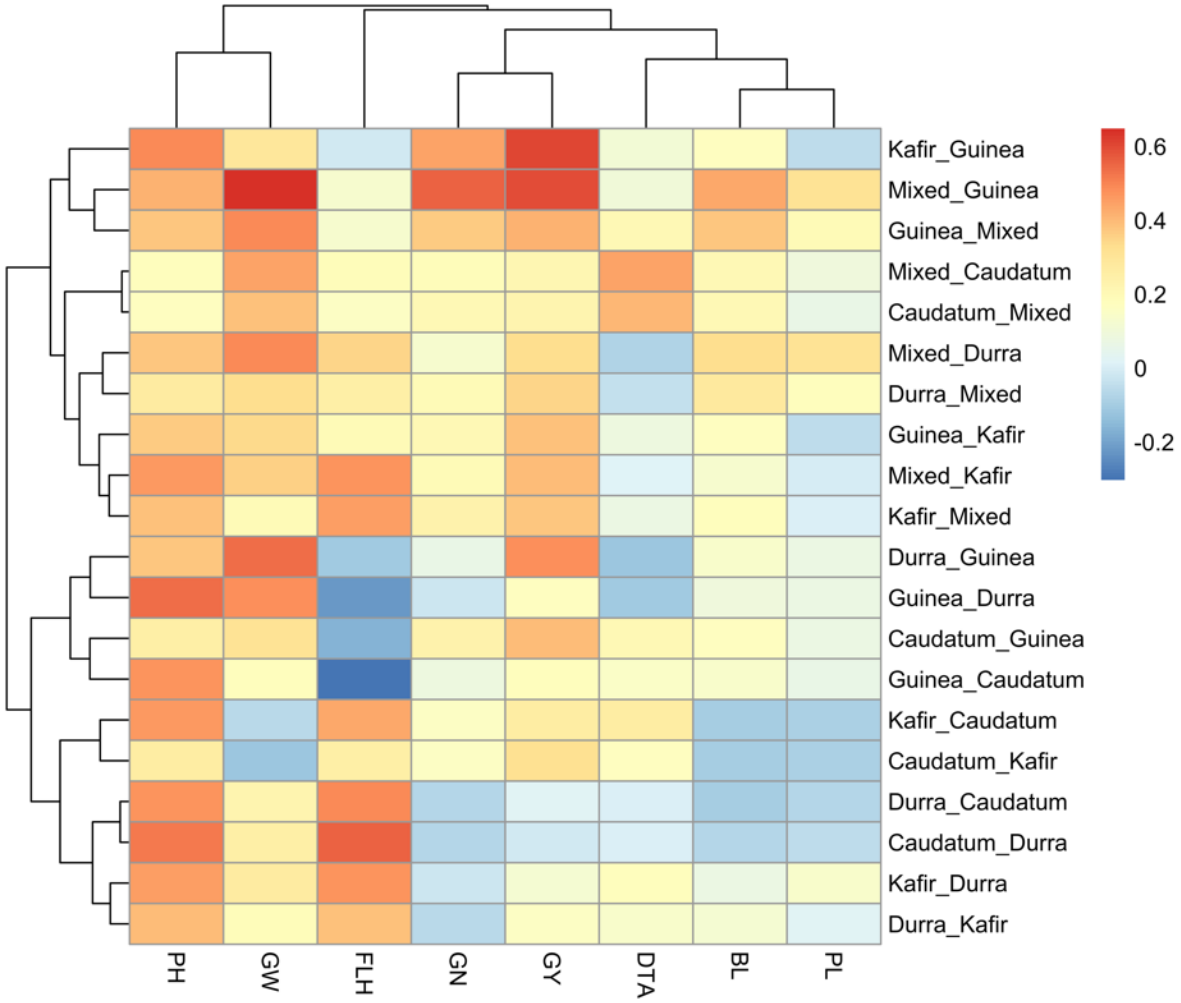

1.00
0.75
0.50
0.25
0.00
h ^ 2
across
within
BL
DTA
FLH
GN
GW
GY
PH
PL

b.


0.75
0.50
0.25
0.00
Covariance
E_race
Cov_race
BL
DTA
FLH
GN
GW
GY
PH
PL

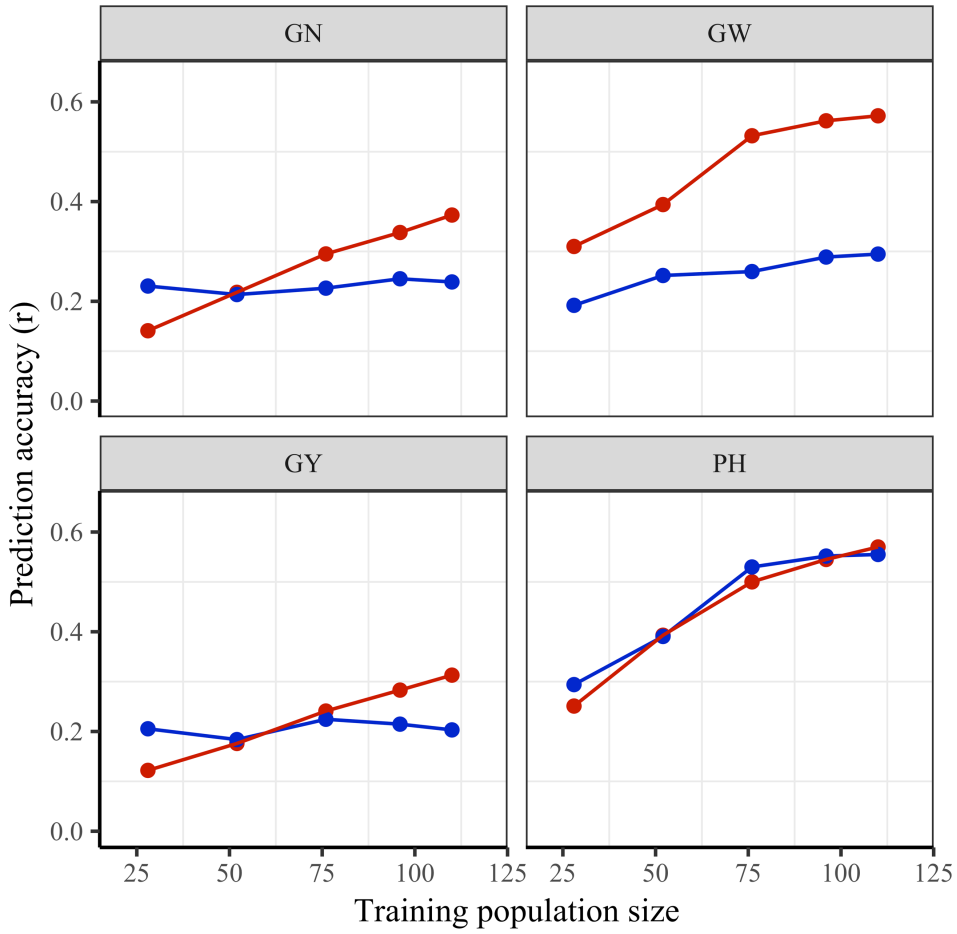
844 Figure 5.

845



846

847 Figure 6.



848