# Genomics-enabled breeding for sorghum improvement in sub-saharan Africa

by

Jacques Martin Faye

B.Sc., University Cheikh Anta Diop of Dakar, 2012
M.S., University Cheikh Anta Diop of Dakar, 2014

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Agronomy
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2020

# Abstract

Sorghum (*Sorghum bicolor*, L. Moench) is a staple cereal food crop for millions of people in Sub-Saharan Africa and Asia. However, drought due to low and unpredictable rainfall decreases its productivity in semiarid regions. Understanding the genetic architecture of adaptive traits (drought tolerance, photoperiodic flowering time, and panicle architecture) of sorghum germplasm from breeding programs across West Africa could contribute to efficient molecular breeding. Breeding priorities in West African sorghum improvement programs seek to develop drought-adapted varieties with yield advantages, early and moderate maturity. However, field phenotyping for adaptation in early generations is difficult and there is limited technology to rapidly develop better-adapted varieties. This study aimed to dissect the genetic architecture of adaptive traits to develop high-throughput breeder-friendly markers for rapid introgression of adaptive alleles from donor to elites lines.

In chapter 1, I describe the sorghum breeding programs in Senegal, the agronomic importance of sorghum types, and genomic approaches for crop improvement in semiarid regions. In chapter 2, I characterize 213,916 single nucleotide polymorphisms (SNPs) across 421 Senegalese sorghum accessions from the USDA-Germplasm Resources Information Network (GRIN) to identify genomic signatures of local adaptation. This study provided insights into the factors shaping the genetic diversity and the molecular systems underlying local adaptation to water scarcity in sorghum, a staple food security crop in Senegal. In chapter 3, I characterize 159,101 SNPs across 756 accessions of the West African sorghum association panel (WASAP) assembled from breeding programs of Senegal, Niger, Mali, and Togo. The genetic diversity structured by botanical types and subpopulations within botanical types across countries and large-effect quantitative trait loci (QTL) for photoperiodic flowering indicate an oligogenic architecture of flowering time in West African sorghum. In chapter 4, I use genome-wide SNP variation from chapter 3 and phenotypic data from multiple managed water stress environments to identify genomic regions associated with drought response. Significantly positive pleiotropic associations contributed to high phenotypic variance and colocalized with known stay-green (*Stg*) QTLs, suggesting the existence of *Stg* alleles in West African sorghum. Finally, in chapter 5, I summarize the expected steps to establish genomics-enabled breeding for sorghum improvement in West Africa.

The genomic resources developed in this research have allowed for the dissection of the genetic architecture of adaptive traits. The SNPs associated with large-effect QTLs can be converted into high-throughput breeder-friendly markers for use in marker-assisted selection. These resources combined with discoveries from the global scientific community can be used to accelerate and facilitate the development of locally adapted varieties to meet global food demand in semiarid regions of Sub-Saharan Africa.

# Genomics-enabled breeding for sorghum improvement in sub-saharan Africa

by

Jacques Martin Faye

B.Sc., University Cheikh Anta Diop of Dakar, 2012
M.S., University Cheikh Anta Diop of Dakar, 2014

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Agronomy
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2020

Approved by:

Major Professor
Geoffrey P. Morris

# Copyright

# Abstract

Sorghum (*Sorghum bicolor*, L. Moench) is a staple cereal food crop for millions of people in Sub-Saharan Africa and Asia. However, drought due to low and unpredictable rainfall decreases its productivity in semiarid regions. Understanding the genetic architecture of adaptive traits (drought tolerance, photoperiodic flowering time, and panicle architecture) of sorghum germplasm from breeding programs across West Africa could contribute to efficient molecular breeding. Breeding priorities in West African sorghum improvement programs seek to develop drought-adapted varieties with yield advantages, early and moderate maturity. However, field phenotyping for adaptation in early generations is difficult and there is limited technology to rapidly develop better-adapted varieties. This study aimed to dissect the genetic architecture of adaptive traits to develop high-throughput breeder-friendly markers for rapid introgression of adaptive alleles from donor to elites lines.

In chapter 1, I describe the sorghum breeding programs in Senegal, the agronomic importance of sorghum types, and genomic approaches for crop improvement in semiarid regions. In chapter 2, I characterize 213,916 single nucleotide polymorphisms (SNPs) across 421 Senegalese sorghum accessions from the USDA-Germplasm Resources Information Network (GRIN) to identify genomic signatures of local adaptation. This study provided insights into the factors shaping the genetic diversity and the molecular systems underlying local adaptation to water scarcity in sorghum, a staple food security crop in Senegal. In chapter 3, I characterize 159,101 SNPs across 756 accessions of the West African sorghum association panel (WASAP) assembled from breeding programs of Senegal, Niger, Mali, and Togo. The genetic diversity structured by botanical types and subpopulations within botanical types across countries and large-effect quantitative trait loci (QTL) for photoperiodic flowering indicate an oligogenic architecture of flowering time in West African sorghum. In chapter 4, I use genome-wide SNP variation from chapter 3 and phenotypic data from multiple managed water stress environments to identify genomic regions associated with drought response. Significantly positive pleiotropic associations contributed to high phenotypic variance and colocalized with known stay-green (*Stg*) QTLs, suggesting the existence of *Stg* alleles in West African sorghum. Finally, in chapter 5, I summarize the expected steps to establish genomics-enabled breeding for sorghum improvement in West Africa.

The genomic resources developed in this research have allowed for the dissection of the genetic architecture of adaptive traits. The SNPs associated with large-effect QTLs can be converted into high-throughput breeder-friendly markers for use in marker-assisted selection. These resources combined with discoveries from the global scientific community can be used to accelerate and facilitate the development of locally adapted varieties to meet global food demand in semiarid regions of Sub-Saharan Africa.

# Table of Contents

# List of Figures

# List of Tables

# Acknowledgements

This research would not have been possible without the support, advice, efforts, and motivation of a wide range of people. I would like to thank my advisor, Dr. Geoffrey Morris for his tremenduous support, mentorship, and always willing to teach me new things throughout the course of this journey. I am grateful for having you as my advisor and for teaching how to be a good scientist and leader. I would like to sincerely thank my committee chair, Dr. Loretta Jonhson and all my committee members, Dr. Robert Aiken, Dr. Guihua Bai, Dr. Eduard Akhunov, and Dr. Ramasamy Perumal for all their support and guidance during my graduate study program. Breeders, geneticists, eco-physiologists, technicians in West Africa, in particular ISRA/CERAAS and Morris lab members have greatly contributed to the success of this research, and for that I am grateful. I would like to thank the Dr. Daniel Fonceka, Dr. Ndiaga Cisse, and Professor Diaga Diouf for their advice and mentorship. Thanks to Dr. Cyril Diatta, Dr. Jack Akata, and Dr. Bassirou Sine for our great interactions and field trips at Bambey and Sinthiou Maleme, Senegal. A special thanks goes to the SMIL staff for all their support and dedecation. I would like to thank the Valentin Family for making Manhattan, Little Aple feels like home to me. Finally, I would like to give special thanks to my family for their unconditional support, patience, and love.

# Dedication

I dedicate this work to all smallholder famers who plant seeds every growing season, often without seeing a drop of rain, to feed their families and the world. As some of Dr. Borlaug's definitions for farmers–"mostly small and humble–who for many years have been fighting a quiet, oftentimes loosing war on the food production front", their hardwork and optimism are sources of inspiration. My hope is to see them produce sustainably enough food to improve their standard of living.

# Chapter 1 - Sorghum Improvement in Semiarid Regions of West Africa

Sorghum (*Sorghum bicolor*, L. Moench) is a subsistence crop for millions of people in Sub-Saharan Africa and Asia (Mundia *et al.*, 2019). Sorghum has several usages depending on the region of production. For instance, while the crop is used for livestock, beverage, and biofuel production in the United States of America and China, it is a staple food crop in Africa and other Asia countries (Emmambux & Taylor, 2013). Among cereal crops, sorghum can adapt to harsh environments where the other cereals such as corn, wheat, and rice cannot survive. In Sub-Saharan Africa, sorghum is grown mostly in areas where annual rainfall fluctuates around 400 mm and 600 mm (Emmambux & Taylor, 2013). Although sorghum is one of the most adapted cereal crops to water deficit, various drought scenarios affect its production. The improvement of this crop for yield stability under drought is crucial to ensure global food security in developing countries of West Africa.

Adverse environmental conditions such as drought and heat can lead to poor crop adaptation and yield losses. Severe pre- and post-flowering water limitation affects pollen development and grain filling, respectively (Barnabás *et al.*, 2008). Understanding of the physiological and genetic basis of factors controlling crops yield stability under drought conditions facilitates the development of well-adapted crop varieties. Yield stability under drought is a complex phenomenon that requires synergic efforts across disciplines, including crop physiology, soil science, breeding, genetics and genomics, and molecular biology. The identification of genes and quantitative trait loci (QTL) associated with grain yield and yield components is crucial for understanding the genetic and molecular basis of drought tolerance. Predictive breeder-friendly high-throughput markers could be designed from candidate loci associated with drought tolerance. Breeders can use these markers to accelerate drought tolerance improvement via marker-assisted selection.

## Production and Importance of Sorghum

Sorghum is a staple food crop for 500 millions of people, mostly living in semi-arid areas in Africa and Asia (Mundia *et al.*, 2019). The word sorghum production is estimated at 63.5 million metric tons in 2015 (Mundia *et al.*, 2019). Africa is the largest sorghum producer in the world with 24.8 million metric tons per year. The US comes at the second position with 17.2

million metric tons, followed by Asia (8.0), and South America (5.5). The overall annual sorghum production has dropped in the last three decades from 62.8 to 59.3 million tons and from 44.5 to 41.9 million hectares due to reduced arable lands and increased harsh environments (FAOSTAT, 2011). These reductions are at a worldwide scale; however environmental effects are more severe at the regional scale because some regions are more affected by environmental changes (Rosenzweig *et al.*, 2014; Mundia *et al.*, 2019). Efforts have been taken by scientific communities to increase its productivity by developing superior resilient varieties. However, improving grain yield under water limitation is challenging in small breeding programs. In Sub-Saharan Africa, post-flowering drought stress is more common due to the early end of rainy seasons. Sorghum breeding programs in West Africa have been targeted to develop varieties with early maturity cycles to escape end-season drought.

## Sorghum Breeding in Senegal

Sorghum improvement started in 1950 (Mauboussin *et al.*, 1977). Mass selection was the first breeding method used by the program to develop pure cultivars from local varieties. Local cultivars are mostly the guinea sorghum type with high grain tannin content, open panicle, tall plant, and low productivity. The first breeding objective of the program was to improve yield performance of local cultivars while maintaining their characteristics similar to farmers' varieties (landraces). From this selection, three main cultivars, 50-59 or Gor-Gatna (caudatum sorghum type, short stature, good performance and stability), 63-18 or Hadien Kori (caudatum sorghum, short stature and early maturity), and SH 60 or Congossane (guinea sorghum, non tannin, late maturity and tall stature) were developed.

After 1965, the breeding objective was to develop varieties that would fit the growing season length in the northern agro-ecological zone (semi-arid zone). Through pedigree selection, several varieties were developed from hybridization between improved cultivars and introduced lines with a maturity cycle varying from 90 to 125 days (Mauboussin *et al.*, 1977). The objective consisted also of creating varieties with short plant stature, semi-compact panicle to reduce pathogen infection such as grain mold, and low tannin content. The two varieties, CE 67 and CE 90 (IRAT 11), which have good agronomic performance, were developed. IRAT 11 was developed from a cross between 63-18 and 63-43 or Mourmoure (an introduced local variety from Niger).

In 1980, the variety IRAT 204 (CE 151-262), locally named 80-25, was created from the cross between IRAT 11 (CE 90) and 73-71 (known as IS 12610). It is a caudatum type with 110 cm of plant height, red plant color, photoperiod insensitive, maturity cycle of 95 days, making it adapted to the semi-arid zone. IRAT 204 was mainly developed particularly for the Soudano-Sahelian agro-climate zone (annual rainfall of 500 to 700 mm). This variety has been used in several breeding programs in West Africa because of its high tolerance to biotic and abiotic stresses such as tolerance to grain mold, drought, and leaf diseases. IRAT 204 is also shown to be resistant to sugarcane aphid (Kebede *et al*., under review). IRAT 204 is a member in the sorghum association panel, which consists of diverse sorghum lines with key agronomic traits, lines used during the sorghum conversion program, and lines from diverse geographic regions (Casa *et al*., 2008). Whole genome de novo sequencing of IRAT 204 has been completed at HudsonAlpha Genome Sequencing Center as part of the Gates Foundation project led by the Donald Danforth Plant Science Center.

The current breeding program of Senegal established in 2000 aimed to develop lines that adapt to the Sahelo-Soudanian and Soudanian agro-climatic zones and to improve grain quality of the varieties developed by the previous programs. The variety CE 180-33 was the recommended variety to grow in these zones because of its its good productivity and tolerance to water-limited conditions. However, its grain quality is low with notable high tannin content, which prevented its expansion in farmers' fields. In 2001, to improve grain quality while maintaining a early maturity, a cross was made between the parents IRAT 204 and Sorvato-1, a line with good grain quality from Togo. From this cross, four varieties, Nguinthe, Faourou, Nganda, and Darou were developed and released in 2011, and two other varieties Payenne (ISRA 618-1) and Golobe (ISRA 618-2) were developed in 2015 using pedigree selection.

The current breeding goal is to develop new varieties with drought tolerance, early and moderate maturity, low tannin content, and semi-compact panicle Grain quality is also a breeding target for these varieties. Nganda is being used as a parent in several breeding crosses. This variety is also the common parent of the nested-association mapping (NAM) population that is being developed at the Centre d'Etude Régional pour l'Amélioration de l'Adaptation à la Sécheresse (CERAAS) in Senegal. The availability of predictive DNA markers for these phenotypes will accelerate the achievement of the current and future breeding goals.

## Genetic Structure and Botanical Characteristics in Cultivated Sorghum

Genetic structure analysis investigates the species organization into populations and factors shaping population structure. Uncovering the genetic ancestry of sorghum accessions according to agro-climatic zones and botanical types helps better design breeding strategies for different regions. Population genomics studies have allowed the characterization of worldwide sorghum germplasm (Deu *et al.*, 2006; Morris *et al.*, 2013). This is relevant for designing core collections, germplasm conservation and breeding targets, identifying parental lines, and accounting for background effects in association mapping studies. Based on genetic distance calculation, heterotic groups can be identified and classified. The genetic variants or alleles that significantly differentiate distinct populations can be used to improve new cultivars for local adaptation. For instance, $F_{ST}$ genetic differentiation scan is one of the most used methods to identify natural variants that differentiate distinct populations.

Among cereal crops, sorghum bicolor is more related to maize (*zea mais*) and rice. Sorghum diverged from maize and rice ~15 and 50 million years ago, respectively. This close relationship indicates the existence of significant conserved genomic sequences, which facilitates comparative genomic analysis (Hamblin *et al.*, 2004; Brown *et al.*, 2006). In the wild relative sorghums, four wild races have been defined, *Arundinaceum*, *Virgatum*, *Aethiopicum*, and *Verticilliflorum*. These races are grouped into *S. bicolor subspecies verticilliflorum* (Harlan & De Wet, 1972). The sorghum wild relatives are not often used in breeding programs as it is the case in other crops such as wheat where wild relatives are usually used to improve for disease resistance. In the cultivated sorghum (*Sorghum bicolor subsp.bicolor* (Linn.) Moench), five basic botanical types and ten intermediate types have been defined based on the spikelet and panicle morphology (Harlan & De Wet, 1972; Deu *et al.*, 1994). These types, including bicolor, durra, guinea, caudatum, and kafir are originally distributed with respect to geographic regions across Africa. These types, except bicolor are used by sorghum breeders at different levels based on their agronomic importance (Fig. 1).

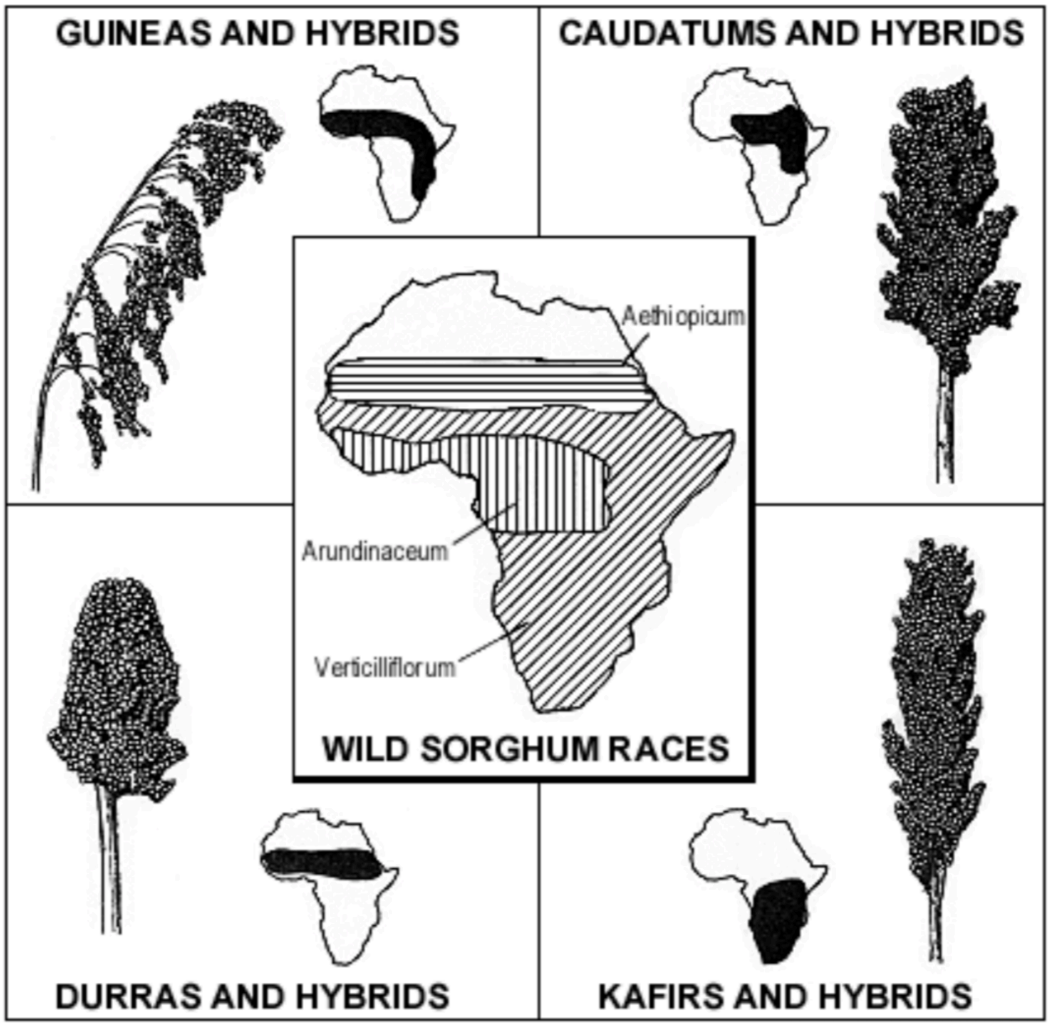



**Figure 1-1. Spikelet morphology of cultivated sorghum types. Courtesy of J Hancock *et al.* (2004) obtained from Harris *et al.* (2007).**

Firstly, bicolor type is thought to be the most primitive among the five basic sorghum types, thus more related to the wild relatives, particularly *verticilliflorum*. The other basic types are thought to be derived from hybridization between bicolor and wild relatives across Africa (Doggett, 1988). Bicolor is the most poorly represented type among cultivated sorghums. It is characterized by loose panicles, small and elongated seeds that are entirely covered by the glumes. Despite being adapted to humid areas, the bicolor race is less productive thus making it undesirable to use for sorghum improvement. However, some bicolor cultivars are grown in small areas as forage and sweet sorghum for the production of syrup.

The guinea type is predominantly distributed in humid savannas of West Africa where it is thought to form a second center of domestication (Folkertsma *et al.*, 2005). Guinea is also present in South-eastern Africa and Asia. It is the most diversified type of cultivated sorghum (Folkertsma *et al.*, 2005). It is characterized by an open panicle, long peduncle, small elliptical seeds that rotate in open and long glumes at maturity. Guinea type contains tall, photoperiod-sensitive, and rustic genotypes highly adapted to humid areas. Several subdivisions exist within the guinea race, *Guinea margaritiferum* which is found in West Africa and has red grain that is used to produce local beer; *Guinea gambicum*, *Guinea guineense*, *Guinea conspicuum* which are found in Eastern Africa and India, and *Guinea roxburghii*. The guinea type is resistant to insects and pathogen infection such as grain mold. The grain texture and presence of tannin in guinea seeds are favorable for the production of local beer in many African countries.

The durra type is predominant in Eastern Africa where it was first domesticated before diffusing to West and Central Africa, Middle East, India, and South Asia (Doggett, 1988). It is thought that durra resulted from introgression of early bicolor with wild forms in dry areas. This type is more adapted to dry areas, especially in the Sahelian zone of Africa. It has a compact panicle, usually a curve peduncle, a big and globose grain that is partially covered by the glumes. Durra type is usually used in breeding to improve sorghum for drought adaptation.

The caudatum type is mainly distributed in Central and in the tropical areas of Eastern Africa. It has a semi-compact panicle with a highly variable panicle shape, generally elliptical. Its grain is lighter with good quality compared to other sorghum types. The grain is inserted in short glumes. The caudatum type is highly appreciated by breeders, especially in the US mainly due to its high yield potential and desired grain texture that is generally free of tannin. These characteristics make caudatum sorghums suitable for hybrid production and inbred line development.

Lastly, the kafir type is found in Southern Africa and is considered as the most recent type. The main characteristic of this type resides from its high yield potential and lack of photosensitivity, thus making it the type of choice of many breeding programs in temperate regions. It has a dense and cylindrical panicle. The glumes are much shorter than the grain. The kafir type was widely used during the sorghum conversion program to develop early maturity and photoperiod insensitive varieties in the US sorghum breeding.

## Improvement for Drought Tolerance

Understanding the physiological and genetic basis of plant water relations contributes to developing drought-resistant varieties (Blum, 2014). Drought tolerance is a complex phenomenon that affects several aspects of the crop (Blum, 2005). Two mechanisms of drought tolerance are known, tolerance to dehydration and avoidance to dehydration. Tolerance to dehydration involves phenological, biochemical, and physiological traits such as osmotic adjustment (e.g. production of proline, glucose, glucane), antioxidant capacity, abscisic acid response, and water soluble carbohydrates reserves (Blum & Arkin, 1984; Blum, 2005, 2014). A study on different wetland species including both monocotyledonous and dicotyledonous plants using water potential, solute potential, and bulk modulus of elasticity found that genotypes used in the experiment showed tolerance to dehydration (Touchette *et al.*, 2007). On the other hand, the avoidance of dehydration occurs mainly through the closure of stomata, which permits plants to maintain sufficient water content (Blum, 2005), change in leaf direction, and deep rooting system.

Managing timing and intensity of the stress is challenging, often semi-controlled screening has limited capacity. Approaches to screen for drought stress involve developing rainout shelters, irrigation systems in low rainfall areas, or the modeling of environmental conditions to reduce abiotic variations for better field control. However, none of these approaches can constrain drought in wide range areas. Water managed stress during off-season allows to better control water deficit in field experiments. However, off-season water stress experiments may not reflect growing seasons, which is the target environment for sorghum farming in smallholder farming.

Heat stress is another factor that limits crop production. Heat stress is usually associated with drought stress (Hatfield & Prueger, 2015). However, heat stress is associated with decreased or deactivation of the photosynthetic capacity, causing oxidative stress and early senescence (Sharkey, 2005; Chen *et al.*, 2012). This leads to a shortened grain filling period. The biochemical aspects such as starch synthesis are also affected by heat stress. Studies have been addressing the physiological aspects of water deficits for a longtime. An integrative approach involves the understanding of the genetic and molecular basis of drought adaptation based on the physiology of drought tolerance (Blum, 2014) to serve breeding programs.

Gene discovery for drought tolerance is crucial for dissecting the genetic basis of yield under drought conditions. To understand the genetic basis of drought tolerance, researches have recently been focusing on the identification of quantitative trait loci (QTLs) and alleles associated with yield components under drought conditions. Genome-wide association studies (GWAS) have been successful in identifying genetic loci associated with agronomic traits of importance in crops (Huang *et al.*, 2012; Morris *et al.*, 2013; Crowell *et al.*, 2016; McCouch *et al.*, 2016; Cao *et al.*, 2016).

The inheritance of yield under drought conditions is complex. Response to drought in sorghum involves delay leave senescence (Kebede *et al.*, 2001). Several drought tolerance QTLs have been reported to contribute to yield under drought and stay-green–delayed leaf senescence under drought conditions (Tuinstra *et al.*, 1997; Xu *et al.*, 2000; Harris *et al.*, 2007; Borrell *et al.*, 2014; Hayes *et al.*, 2016). These stay-green QTLs (*Stg1–4*) contribute between 8 to 30% of phenotypic variance of stay-green. *Stg1* and *Stg2* are located on chromosome 3, *Stg3a* and *Stg3b* are located on chromosome 2, and *Stg4* is located on chromosome 5 (Fig. 1-2) (Mace *et al*., 2019). *Stg1–4* QTLs have been introgressed into elite backgrounds of breeding programs in Australia and India. The known founders of stay-green alleles are BTx642 (formerly known as B35), QL41, E36-1, and SC56. These lines are derived from durra sorghum in Ethiopia. However, it is not known whether the stay-green alleles exist in other Africa sorghum across the Sahelian zone. It is also shown that stay-green alleles increase grain yield by modifying canopy architecture and water supply in lines with introgressed B35 alleles (Borrell *et al.*, 2014). To shed lights on the existence of *Stg1–4* alleles in the West Africa sorghum, chapter 4 describes the phenotypic variation across multiple water stress environments to dissect the genetic architecture of drought tolerance.

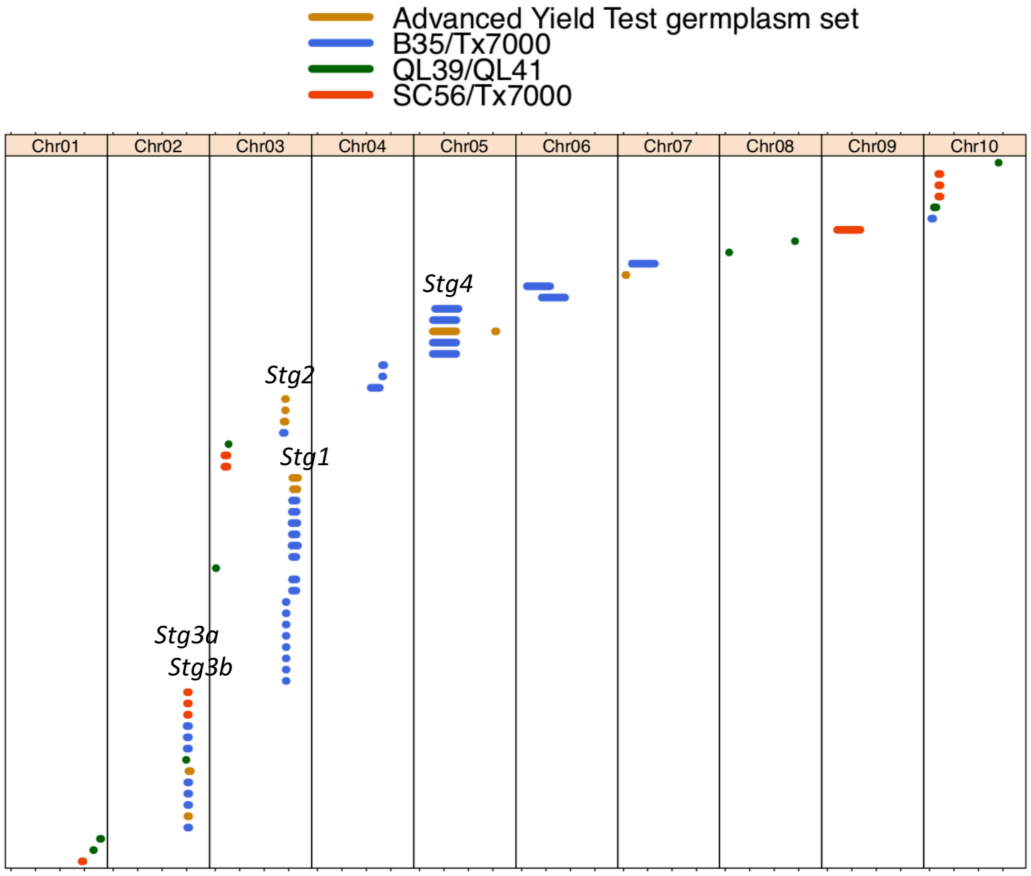



**Figure 1-2. Genomic position of sorghum stay-green quantitative loci from the sorghum QTL Atlas.**

## Genomic Tools for Crop Improvement

GWAS, also known as linkage disequilibrium (LD) mapping, consists of integrating phenotypes and genotypes and use linear models to find association between SNP markers and causative variants of a trait based on LD. This approach was first developed in human genetics to identify common disease-associated variants (Hirschhorn & Daly, 2005). The principle is based on the hypothesis that common variation in phenotype is associated with common genetic variants–common-disease common-variant hypothesis (Manolio *et al.*, 2009). GWAS take advantage of ancient recombinations and genetic diversity that exist within adapted cultivars to identify gene/QTLs. GWAS can be applied on breeding populations and unrelated individuals from different geographic regions. Using numerous tagging SNPs, GWAS allow the dissection of genetic architecture of complex traits by providing the number of loci controlling a trait, their effect size, their frequencies, and gene action that is involved. Because many mutated genes have

been identified from mutation studies, conserved genes can be identified based on existing knowledge from other model systems.

The power of GWAS results from the presence of common variants across a large population sample size. However, this power can be influenced by factors such as the extent of linkage disequilibrium, number of markers, low heritability of trait, presence of rare variants with a small effect on trait, and the confounding effect due to background effects.

Linkage disequilibrium, which is the nonrandom association of alleles at two or more loci on the same chromosome, is crucial for effective GWAS. LD decay provides estimation of the number of SNP markers that is adequate for high-resolution GWAS. The mating system plays a key role in the LD estimation for GWAS. Self-pollinated species usually have a long range of LD compared to outcrossing species. High-density SNP markers are required for high-resolution GWAS. Advances in next generation sequencing technologies (Thomson, 2014) and high-throughput genotyping platforms such as genotyping-by-sequencing (Elshire *et al.*, 2011; Poland *et al.*, 2012) can generate high-density SNP markers at less cost.

Heritability provides insights into the correlation of phenotype to additive genetic effect in association mapping studies. A null heritability would suggest that phenotypic variance is not due to a genetic effect, and therefore carrying out GWAS analysis might be not useful (Korte & Farlow, 2013). Low heritability in GWAS may result in difficulty to identify rare variants. Most of the important traits in crops have low heritability, making GWAS challenging to detect common variants. The problem associated with low heritability can be dealt with by increasing population sample size and phenotypes from multiple environments. In GWAS linear-mixed models, heritability is estimated as pseudo-heritability, which is in most cases lower than the trait heritability. This low heritability is due to missing genome-wide variants that are non-genotyped–missing heritability (Korte & Farlow, 2013). A way to deal with missing heritability would be to use whole genome sequencing (Manolio *et al.*, 2009) or at least to impute missing genotypes to increase the frequency of rare alleles.

Most of the rare variants in GWAS experiments create synthetic associations or indirect associations (Dickson *et al.*, 2010). It is difficult to distinguish these rare variants from true trait-associated SNPs. Most complex traits are controlled by a large number of rare causal alleles that tend to be clustered together in large population size.

Conventional breeding has yielded many elite varieties of important economic values. However, phenotypic selections for many adaptive traits such as photoperiodic flowering and drought tolerance is difficult for small breeding programs because it requires multiple field trials. Understanding the genetic architecture of complex traits contributes to facilitating and accelerating breeding. Favorable trait-associated alleles can be converted into breeder-friendly markers that tag these alleles. These makers can be used in marker-assisted selection (MAS) to rapidly develop adapted varieties (Hash *et al.*, 2003; Varshney *et al.*, 2013).
Genomic selection (GS) has shown to increase genetic gain and reduce the number of breeding cycles. Trait-associated alleles also can be integrated into genomic selection models to increase prediction accuracy (Fu *et al.*, 2017). In GS, the whole performance of individuals is assessed regardless of interaction between positive and negative alleles. GS model estimates the individual total performance. This makes GS a powerful method to detect best performing lines in the population. However, a few to several thousands of markers are used to estimate breeding values of individuals compared to MAS.

The development of genomic tools contributes to solving the roadblocks that limit or delay the process of achieving breeding goals. Both genomic selection and MAS provide advantages to more rapidly increase genetic gain per unit time and unit cycle, thus can be integrated together in the breeding pipeline for more efficient crop breeding. In this dissertation, chapter 1 describes the sorghum breeding program in Senegal, the agronomic importance of sorghum types, and genomic approaches for crop improvement in semiarid regions. To identify genomic signatures of adaptation to semi-arid versus sub-humid climates, chapter 2 describes the genetic characterization of genome-wide nucleotide polymorphisms across the Senegalese sorghum accessions in GRIN. To determine the population structure of the West African sorghum and demonstrate the effectiveness of GWAS analysis to identify known photoperiodic flowering time loci, chapter 3 describes the genetic characterization of the West African sorghum association panel (WASAP) assembled from breeding programs of Senegal, Niger, Mali, and Togo. To develop locally-adapted varieties and rapidly introgress favorable trait-associated alleles from West African donor lines to farmers' preferred varieties, chapter 5 describes the genomic-enabled breeding approaches using the goal-directed hypothesis driven research (GoHy).

## References


Blum A. 2005. Drought resistance, water-use efficiency, and yield potential—are they compatible, dissonant, or mutually exclusive? *Australian Journal of Agricultural Research* 56: 1159.

Blum A. 2014. Genomics for drought resistance – getting down to earth. *Functional Plant Biology* 41.

Blum A, Arkin GF. 1984. Sorghum root growth and water-use as affected by water supply and growth duration. *Field Crops Research* 9: 131–142.

Borrell AK, van Oosterom EJ, Mullet JE, George-Jaeggli B, Jordan DR, Klein PE, Hammer GL. 2014. Stay-green alleles individually enhance grain yield in sorghum under drought by modifying canopy development and water uptake patterns. *New Phytologist* 203: 817–830.

Brown PJ, Klein PE, Bortiri E, Acharya CB, Rooney WL, Kresovich S. 2006. Inheritance of inflorescence architecture in sorghum. *Theoretical and Applied Genetics* 113: 931–942.

Cao K, Zhou Z, Wang Q, Guo J, Zhao P, Zhu G, Fang W, Chen C, Wang X, Wang X, *et al.*2016. Genome-wide association study of 12 agronomic traits in peach. *Nature Communications* 7: 13246.

Cavanagh C, Morell M, Mackay I, Powell W. 2008. From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. *Current Opinion in Plant Biology* 11: 215–221.

Chen WR, Zheng JS, Li YQ, Guo WD. 2012. Effects of high temperature on photosynthesis, chlorophyll fluorescence, chloroplast ultrastructure, and antioxidant activities in fingered citron. *Russian Journal of Plant Physiology* 59: 732–740.

Crowell S, Korniliev P, Falcão A, Ismail A, Gregorio G, Mezey J, McCouch S. 2016. Genome-wide association and high-resolution phenotyping link Oryza sativa panicle traits to numerous trait-specific QTL clusters. *Nature Communications* 7.

Deu M, Gonzalez-de-Leon D, Glaszmann J-C, Degremont I, Chantereau J, Lanaud C, Hamon P. 1994. RFLP diversity in cultivated sorghum in relation to racial differentiation. *Theoretical and Applied Genetics* 88: 838–844.

Deu M, Rattunde F, Chantereau J. 2006. A global view of genetic diversity in cultivated sorghums using a core collection. *Genome* 49: 168–180.

Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. 2010. Rare Variants Create Synthetic Genome-Wide Associations. *PLOS Biology* 8: e1000294.

Doggett H. 1988. *Sorghum*. Longman Scientific & Technical.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species (L Orban, Ed.). *PLoS ONE* 6: e19379.

Emmambux MN, Taylor JRN. 2013. Morphology, physical, chemical, and functional properties of starches from cereals, legumes, and tubers cultivated in Africa: A review. *Starch - Stärke* 65: 715–729.

Folkertsma RT, Rattunde HFW, Chandra S, Raju GS, Hash CT. 2005. The pattern of genetic diversity of Guinea-race Sorghum bicolor (L.) Moench landraces as revealed with SSR markers. *Theoretical and Applied Genetics* 111: 399–409.

Fu Y-B, Yang M-H, Zeng F, Biligetu B. 2017. Searching for an Accurate Marker-Based Prediction of an Individual Quantitative Trait in Molecular Plant Breeding. *Frontiers in Plant Science* 8.

Hamblin MT, Mitchell SE, White GM, Gallego J, Kukatla R, Wing RA, Paterson AH, Kresovich S. 2004. Comparative population genetics of the panicoid grasses: sequence polymorphism, linkage disequilibrium and selection in a diverse sample of sorghum bicolor. *Genetics* 167: 471–483.

Harlan JR, De Wet JJM. 1972. A Simplified Classification of Cultivated Sorghum1. *Crop Science* 12: 172–176.

Harris K, Subudhi PK, Borrell A, Jordan D, Rosenow D, Nguyen H, Klein P, Klein R, Mullet J. 2007. Sorghum stay-green QTL individually reduce post-flowering drought-induced leaf senescence. *Journal of Experimental Botany* 58: 327–338.

Hash CT, Bhasker Raj AG, Lindup S, Sharma A, Beniwal CR, Folkertsma RT, Mahalakshmi V, Zerbini E, Blümmel M. 2003. Opportunities for marker-assisted selection (MAS) to improve the feed quality of crop residues in pearl millet and sorghum. *Field Crops Research* 84: 79–88.

Hatfield JL, Prueger JH. 2015. Temperature extremes: Effect on plant growth and development. *Weather and Climate Extremes* 10: 4–10.

Haussmann B, Mahalakshmi V, Reddy B, Seetharama N, Hash C, Geiger H. 2002. QTL mapping of stay-green in two sorghum recombinant inbred populations. *Theoretical and Applied Genetics* 106: 133–142.

Hayes CM, Weers BD, Thakran M, Burow G, Xin Z, Emendack Y, Burke JJ, Rooney WL, Mullet JE. 2016. Discovery of a Dhurrin QTL in Sorghum: Co-localization of Dhurrin Biosynthesis and a Novel Stay-green QTL. *Crop Science* 56: 104–112.

Hirschhorn JN, Daly MJ. 2005. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* 6: 95–108.

Huang X, Zhao Y, Wei X, Li C, Wang A, Zhao Q, Li W, Guo Y, Deng L, Zhu C, *et al.*2012. Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nature Genetics* 44: 32–39.

Huang BE, Verbyla KL, Verbyla AP, Raghavan C, Singh VK, Gaur P, Leung H, Varshney RK, Cavanagh CR. 2015. MAGIC populations in crops: current status and future prospects. *Theoretical and Applied Genetics* 128: 999–1017.

Kebede H, Subudhi PK, Rosenow DT, Nguyen HT. 2001. Quantitative trait loci influencing drought tolerance in grain sorghum (Sorghum bicolor L. Moench). *Theoretical and Applied Genetics* 103: 266–276.

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

Mace, E., Innes, D., Hunt, C., Wang, X., Tao, Y., Baxter, J., Hassall, M., Hathorn, A., & Jordan, D. (2019). The Sorghum QTL Atlas: A powerful tool for trait dissection, comparative genomics and crop improvement. *Theoretical and Applied Genetics*, *132*(3), 751–766. https://doi.org/10.1007/s00122-018-3212-5.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, *et al.*2009. Finding the missing heritability of complex diseases. *Nature* 461: 747–753.

Mauboussin J-C, Gueye J, N'Diaye M. 1977. L'amélioration du Sorgho au Sénégal. *AGRONOMIE TROPICALE XXXll*: 8.

McCouch SR, Wright MH, Tung C-W, Maron LG, McNally KL, Fitzgerald M, Singh N, DeClerck G, Agosto-Perez F, Korniliev P, *et al.*2016. Open access resources for genome-wide association mapping in rice. *Nature Communications* 7: 1–14.

Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, Riera-Lizarazu O, Brown PJ, Acharya CB, Mitchell SE, *et al.*2013. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proceedings of the National Academy of Sciences of the United States of America* 110: 453–458.

Mundia CW, Secchi S, Akamani K, Wang G. 2019. A Regional Comparison of Factors Affecting Global Sorghum Production: The Case of North America, Asia and Africa's Sahel. *Sustainability* 11: 2135.

Poland JA, Brown PJ, Sorrells ME, Jannink J-L. 2012. Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. *PLOS ONE* 7: e32253.

Rosenzweig C, Elliott J, Deryng D, Ruane AC, Müller C, Arneth A, Boote KJ, Folberth C, Glotter M, Khabarov N, *et al.*2014. Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison. *Proceedings of the National Academy of Sciences* 111: 3268–3273.

Sharkey TD. 2005. Effects of moderate heat stress on photosynthesis: importance of thylakoid reactions, rubisco deactivation, reactive oxygen species, and thermotolerance provided by isoprene. *Plant, Cell & Environment* 28: 269–277.

Thomson MJ. 2014. High-Throughput SNP Genotyping to Accelerate Crop Improvement. *Plant Breeding and Biotechnology* 2: 195–212.

Touchette BW, Iannacone LR, Turner GE, Frank AR. 2007. Drought tolerance versus drought avoidance: A comparison of plant-water relations in herbaceous wetland plants subjected to water withdrawal and repletion. *Wetlands* 27: 656–667.

Tuinstra MR, Grote EM, Goldsbrough PB, Ejeta G. 1997. Genetic analysis of post-flowering drought tolerance and components of grain development in Sorghum bicolor (L.) Moench. *Molecular Breeding* 3: 439–448.

Varshney RK, Mohan SM, Gaur PM, Gangarao NVPR, Pandey MK, Bohra A, Sawargaonkar SL, Chitikineni A, Kimurto PK, Janila P, *et al.*2013. Achievements and prospects of genomics-assisted breeding in three legume crops of the semi-arid tropics. *Biotechnology Advances* 31: 1120–1134.

Xu W, Rosenow DT, Nguyen HT. 2000. Stay green trait in grain sorghum: relationship between visual rating and leaf chlorophyll concentration. *Plant Breeding* 119: 365–367.

# Chapter 2 - Genomic Signatures of Adaptation to Sahelian and Soudanian Climates in Sorghum Landraces of Senegal

This chapter has been published as following journal article:
Jacques M. Faye, Fanna Maina, Zhenbin Hu, Daniel Fonceka, Ndiaga Cisse, Geoffrey P. Morris. Genomic Signatures of Adaptation to Sahelian and Soudanian Climates in Sorghum Landraces of Senegal. Ecology and Evolution. DOI:10.1002/ece3.5187

## Abstract

Uncovering the genomic basis of climate adaptation in traditional crop varieties can provide insight into plant evolution and facilitate breeding for climate resilience. In the African cereal sorghum (*Sorghum bicolor* L. [Moench]), the genomic basis of adaptation to the semiarid Sahelian zone versus the sub-humid Soudanian zone is largely unknown. To address this issue we characterized a large panel of 421 georeferenced sorghum landrace accessions from Senegal and adjacent locations at 213,916 single nucleotide polymorphisms (SNPs) using genotyping-by-sequencing. Seven subpopulations distributed along the north-south precipitation gradient were identified. Redundancy analysis found that climate variables explained up to 8% of SNP variation, with climate collinear with space explaining most of this variation (6%). Genome scans of nucleotide diversity suggest positive selection on chromosome 2, 4, 5, 7, and 10 in durra sorghums, with successive adaptation during diffusion along the Sahel. Putative selective sweeps were identified, several of which co-localize with stay-green drought tolerance (*Stg*) loci, and *a priori* candidate genes for photoperiodic flowering and inflorescence morphology. Genome-wide association studies of photoperiod sensitivity and panicle compactness identified 35 and 13 associations that co-localize with *a priori* candidate genes, respectively. Climate-associated SNPs co-localize with *Stg3a*, *Stg1*, *Stg2*, and *Ma6* and have allelic distribution consistent with adaptation across Sahelian and Soudanian zones. Taken together, the findings suggest an oligogenic basis of adaptation to Sahelian versus Soudanian climates, underpinned by variation in conserved floral regulatory pathways and other systems that are less understood in cereals.

## Introduction

Local adaptation is critical for survival of traditional crop varieties in stressful environments (Camus-Kulandaivelu et al., 2006; Xu et al., 2006). Smallholder farmers in

developing countries are particularly vulnerable to environmental factors such as drought and heat stress limiting crop production (Morton, 2007). Climatic gradients in relation with precipitation are major drivers of adaptation in plants including traditional crop varieties (Fournier-Level et al., 2011; Lasky et al., 2015, 2012; Siepielski et al., 2017; Vigouroux et al., 2011). Adaptation to water-limited environments involves phenological, physiological, and morphological traits such as photoperiod sensitivity, delayed senescence, and inflorescence morphology (Blum, 2014). For instance, when growing seasons are shortened by end-of-season droughts, selection favors early maturity alleles to escape drought (Franks et al., 2007; Kenney et al., 2014; Vigouroux et al., 2011). Identifying genetic polymorphisms underlying adaptive traits and their eco-geographic distributions is necessary to understand the genetic basis of local adaptation of landraces (Romero Navarro et al., 2017).

The patterns of genome-wide nucleotide polymorphisms provide insight into selective forces varying over time and space (Olsen et al., 2006; Slatkin, 2008). Recent studies in rice (Caicedo et al., 2007; Li et al., 2017), tomato (Lin et al., 2014), and maize (Swarts et al., 2017) have shown that high genetic differentiation among populations reflects adaptation to specific agroclimatic zones. Population genomic approaches for identifying signatures of selection include decreased pairwise nucleotide diversity, composite likelihood ratio (CLR) analysis for selective sweeps, and genome-environment associations (Fang et al., 2017; Fournier-Level et al., 2011; Lasky et al., 2015; Li et al., 2017; Lin et al., 2014). The CLR analysis in SweeD is relatively robust to demographic events because the method conservatively estimates the neutral site frequency spectrum (SFS) based on the observed data (Nielsen et al., 2005; Pavlidis et al., 2013). Linear regression models and genome-wide association studies (GWAS) mixed models are common methods used for GEA, especially to investigate adaptation to environmental gradients (Rellstab et al., 2015) and have been applied by several studies in plants and crop species (Fournier-Level et al., 2011; Lasky et al., 2015; Yoder et al., 2014). Redundancy analysis (RDA) provides an estimate of allelic variance explained by climatic factors based on multivariate linear regressions (Meirmans, 2015). Genome-wide association studies can provide high mapping resolution of adaptive traits in diverse populations (Cavanagh et al., 2008).

Sorghum (*Sorghum bicolor* L. [Moench]) is a staple food crop for smallholder farmers in semi-arid regions worldwide. The modest genome size (~800 Mbp) of sorghum relative to other grass species (Paterson et al., 2009) makes it a tractable system for the genomic studies of local

adaptation. Five botanical types (bicolor, durra, guinea, caudatum, and kafir) have been described (Harlan and De Wet, 1972). Durra types, known for their adaptation to arid zones, are thought to have originated in Ethiopia before westward diffusion along the Sahel to West Africa and finally Senegal (Harlan and De Wet, 1972). Guinea types, known for their humid adaptation (Deu et al., 1994; Folkertsma et al., 2005), may reflect a second center of domestication in the humid savanna of West Africa (Deu et al., 1994; Doggett, 1988; Folkertsma et al., 2005). Inflorescence morphology is a major component of agroclimatic adaptation in sorghum and varies from loose panicle in guinea to compact panicle in durra sorghum (Brown et al., 2006). Most traditional sorghum varieties in West Africa are photoperiod sensitive such that grain maturation coincides with the end of the rainy season (Bhosale et al., 2012; Sanon et al., 2014). In U.S. sorghum, variation in flowering time is controlled by conserved cereal floral regulatory networks, including phytochromes (*Ma3/PhyB*, *Ma5/PhyC*), CCT-domain regulators (*Ma1/PRR37*, *SbEhd1, SbEhd2)*, and florigens (*SbCN15*/*Hd3a*, *SbCN12*) (Mullet et al., 2010; Murphy et al., 2011). Several quantitative trait loci (QTL) (*Stg1–4*) confer stay-green (i.e. delayed leaf senescence) post-flowering drought tolerance in lines derived from Ethiopian durra (Borrell et al., 2014; Harris et al., 2007; Kebede et al., 2001; Tuinstra et al., 1997), but it is not known whether these loci contribute to drought adaptation more widely across the Sahel.

Analyses of genetic diversity, linkage disequilibrium (LD), and genome-environment associations have provided an understanding of worldwide sorghum genetic structure across diverse agroclimatic regions (Bouchet et al., 2012; Lasky et al., 2015; Mace et al., 2013; Morris et al., 2013; Wang et al., 2013). However, the genomic basis of climate adaptation at a regional scale remains poorly understood. The variation of agroclimatic conditions in Senegal reflects the sub-Saharan climatic gradient with increasing annual precipitation from north to south across the Sahelian (~200–600 mm) and Soudanian zones (~600–1100 mm). A large panel of sorghum landraces was collected from these agroclimatic zones in Senegal in the 1970s (Clément and Houdiard, 1977). To better understand the genomic basis of Sahelian and Soudanian climate adaptation, we used genotyping-by-sequencing (GBS) to characterize genome-wide single nucleotide polymorphism (SNP) in georeferenced and phenotyped Senegalese sorghum landraces. We characterized population structure of genomic diversity, identified signatures of selection, and mapped genetic polymorphisms associated with phenotype and climate. The

findings suggest that climate has shaped genomic variation across Sahelian and Soudanian zones, with variation in floral regulatory pathways and other systems contributing to this adaptation.

## Materials and Methods

### Plant materials

The Senegalese sorghum germplasm (SSG) used in the present study were obtained from the U.S. Department of Agriculture (USDA) Germplasm Resources Information Network (GRIN). These accessions (n = 341) were collected from various agro-ecological zones of Senegal in 1976 (Clément and Houdiard, 1977). GRIN accessions from neighboring countries of Gambia (n = 60), which is surrounded by Senegal, and Mauritania (n = 15), which shares border along the Senegal River Valley, were also included in our panel. Six improved varieties (CE 151-262, CE 180-33, ISRA-S-621-B, 53-49, CE 260-12-1-1, IRAT 4) from the sorghum breeding program based at the Centre National de Recherche Agricole (CNRA) and two sorghum conversion lines, SC 1067 (PI 576432) and SC 417 (PI 533861), were included. Information about the SSG including botanical race, geographic origin, local name, and ethno-linguistic group from which the landrace was collected are presented in Supporting Information Data S1. Assignment in "durra" group was from the GRIN genebank, based on a phenotypic assessment. To compare the SSG landraces with the global sorghum diversity, we reanalyzed available raw sequencing data of worldwide sorghum diversity panels (Morris et al., 2013), hereafter referred to as the global diversity panel (GDP). This data set included 582 lines from the sorghum mini core collection and the Generation Challenge Program reference set, and 178 lines from the sorghum association panel. The GDP includes accessions from Africa, Asia, and the Americas.

### Genotyping-by-sequencing

Accessions of the SSG were grown in a greenhouse at Kansas State University. Leaf tissues from each accession were harvested from two weeks old seedlings (five seedlings pooled per accession), placed into 96-well plates, and dried in a lyophilizer for two days. Genomic DNA of SSG accessions was extracted from ~50 mg dried leaf tissue using the BioSprint robot with DNeasy Mini Kit (Qiagen) according to the manufacturer's instructions. DNA was quantified with PicoGreen and normalized to 10 ng/µl DNA for each sample. The GBS library was constructed using the restriction enzyme *Ape*KI for DNA digestion and 384-plex barcode ligation (4 x 96-plex) following the GBS protocol (Elshire et al., 2011). Digested DNA fragments were

ligated to the barcode-adapters in a solution containing the 10x T4 DNA Ligase Reaction Buffer, ultra-pure water, and T4 DNA Ligase (New England Biolabs), then cleaned using a QIAquick PCR purification kit (Qiagen). The adapter-ligated DNA fragments were amplified by polymerase chain reaction (PCR). The PCR-amplified DNA fragments were cleaned and quantified with PicoGreen. Four 96-plex libraries were pooled to form a 384-plex GBS library. GBS libraries were diluted into 20 µl at 4 nM for each library and analyzed by the Agilent 2100 Bio-analyzer for sequencing. GBS libraries were sequenced on Illumina HiSeq 2500 at the University of Kansas Medical Center.

**SNP calling**

The SNP calling was done based on 1208 samples including the accessions from the SSG panel and accessions from the GDP. Single-end sequence reads obtained from Illumina sequencing and raw sequencing data from the GDP were processed with the TASSEL 5 GBS v2 pipeline (Glaubitz et al., 2014). All unique sequence reads were trimmed to 64 bp, which was the default setting. The first step in the pipeline (GBSSeqToTagDBPlugin) allowed to collapse identical reads into tags using the key files of both SSG and GDP accessions. Distinct tags were pulled and exported from the database in the fastQ format using the TagExportToFastqPlugin for their alignment to the BTx623 sorghum reference genome v.3.1 (McCormick et al., 2018; Paterson et al., 2009). The alignment was performed with the Burrows-Wheeler Alignment (Li and Durbin, 2009) where the created SAM file was passed through the SAMToGBSdbPlugin to store the position information of aligned tags. The SNPs were called from the aligned tags. The DiscoverySNPCallerPlugin was used to identify SNPs from the aligned tags where MAF was set to 0.0001 and minimum locus coverage (mnLCov) was kept as the default setting of 0.1. For downstream population genomic analyses, SNPs with < 20% missing data rate and minor allele frequencies (MAF) > 0.01 were retained. Monomorphic sites were removed and only biallelic sites were retained. Missing genotypes were imputed using Beagle v4.1 program (Browning & Browning, 2016). For the association mapping studies, the SNP dataset was filtered for MAF > 0.05 to reduce the chance of observing false positive associations.

**Population structure analysis**

Principal component analysis (PCA) of SNP variation was performed using the snpgdsPCA function of the R package SNPRelate (Zheng et al., 2012). Neighbor-joining analysis was performed using TASSEL 5 program and the tree was visualized with the ape

package in R (Paradis et al., 2004). Bayesian model-based clustering in ADMIXTURE v1.23 (Alexander et al., 2009) was used to estimate the subpopulation membership/admixture for K = 2–20 subpopulations. To reduce SNP redundancy due to LD for the admixture analysis, genotypic data was LD-pruned with a *window size* of 50 SNPs, *step size* 10, and *VIF threshold* of 0.5 using the function *indep* in PLINK 1.9 (Purcell et al., 2007). Default settings of ADMIXTURE were used and five-fold cross validation (CV) error with block bootstrap and 2000 iterations was used to determine the optimum value of K. Each accession was assigned to subpopulation when the proportion of the coefficient of membership to subpopulation was greater than 0.60. To determine the spatial genetic co-ancestry structure with respect to geography, we used the R package TESS3 (Caye et al., 2016). Results were visualized using the R program (R Core Team, 2016).

## Linkage disequilibrium analysis

LD was characterized in the whole SSG and separately in the guinea and durra accessions. VCFtools (Danecek et al., 2011) was used to filter the genotypic data based on MAF > 0.05. The pairwise correlation coefficient ($r^2$) among SNPs was used to estimate LD using TASSEL 5 (Bradbury et al., 2007). LD decay, measured as the distance by which the $r^2$ decays to half its maximum value, was fit using the nonlinear least square (*nls*) function (Hill and Weir, 1988; Remington et al., 2001) in R program. The R package LDheatmap 0.99-4 (Shin et al., 2006) was used to determine and display the pairwise LD surrounding (50 kb region from both sides of the SNP) a SNP-environment variable association.

## Genome-wide nucleotide variation and genome scans

MAF and observed and expected heterozygosity for SNP markers were calculated using VCFtools program and R program (R Core Team, 2016). Pairwise genetic differentiation ($F_{ST}$) among subgroups defined based on eco-geography was estimated using the Weir and Cockerham method in VCFtools. $F_{ST}$ values among subgroups obtained at K = 7 from the TESS3 program were calculated using the R package HierFstat (de Meeûs and Goudet, 2007). Pairwise genome-wide nucleotide diversity ($\pi$) and Tajima's *D* test statistics were calculated based on non-overlapping sliding windows of 1 Mbp across the genome using VCFtools. Ratios of $\pi$ were analyzed between guinea and durra accessions in the SSG ($\pi_{guinea} / \pi_{durra}$), and across putative pre-bottleneck and post-bottleneck events ($\pi_{guinea} / \pi_{Ethiopia\ durra}$, $\pi_{Ethiopia\ durra} / \pi_{Niger\ and\ Mali\ durra}$, and $\pi_{Niger\ and\ Mali\ durra} / \pi_{Senegal\ durra}$). Selective sweeps were detected using the composite likelihood

ratio (CLR) method in SweeD program (Pavlidis et al., 2013). Each chromosome was divided into 5,000 grid points (non-overlapping windows). The CLR windows with ≥ 8 SNPs (approximately 1 SNP per 2 kb) were retained during the analysis. The significance threshold representing the 95$^{th}$ percentile cutoff was determined based on 1,000 simulations.

## Genome-wide association studies (GWAS)

GWAS were carried out using mixed-linear models (MLM) in GAPIT in R (Lipka et al., 2012) with the three first principal components eigenvectors and kinship matrix. The Bonferroni correction at $\alpha = 0.05$ level was used to define the significance of association tests. SNPs were filtered at MAF > 0.05, yielding 145,235 SNPs. Phenotypic data was obtained from the GRIN database and treated as binary data for both photoperiod sensitivity (e.g. sensitive versus insensitive) and panicle compactness (e.g. compact versus open panicle). For genome-environment associations (GEA), both MLM and general linear models (GLM) were used. Nineteen WorldClim-derived bioclimatic variables (Hijmans et al., 2005) were used for genome-environment association tests. To identify environment-associated SNPs with the greatest significance among SNPs of the same genomic region, the multi-locus mixed-model (MLMM) (Segura et al., 2012) was used to complement the GLM and MLM. In both MLM and MLMM, the first three principal components were included to account for population structure.

## A priori candidate genes

A list of *a priori* candidate genes for climate adaptation was defined from known sorghum genes, orthologs of cloned genes from rice and maize, and candidates from previous sorghum mapping studies (see Supporting Information Data S2 for candidate genes, gene functions, and references). A literature survey of sorghum orthologs of maize and rice genes that affect inflorescence architecture, flowering time, and drought tolerance was carried out. Inflorescence architecture candidate genes from a previous global GWAS (Morris et al., 2013), photoperiodic flowering time candidate genes from a previous study (Bhosale et al., 2012), and validated drought tolerance loci (stay-green, *Stg1–4*) from (Borrell et al., 2014) were included. Genomic position of candidate genes were determined using Phytozome v12.1.6 (https://phytozome.jgi.doe.gov) (Goodstein et al., 2012).

## Redundancy analysis

RDA was performed using the R package vegan (Oksanen et al., 2017) for climatic factors, ethnicity, and space. Independent variables included 19 climatic variables, space

variables (latitude and longitude), and ethnicity variables. Ethnicity was coded as binary variable indicating the ethno-linguistic group of the farmer that contributed the landrace to the collection (Clément and Houdiard, 1977). Forward selection based on 1,000 permutations was performed for space (e.g., using polynomial coordinates), climate, and ethnicity variables to include only the meaningful variables for ordination. The total among-population genetic variance was partitioned into space, climate, ethnicity, and their overlapping fractions using 1,000 randomly selected SNP (MAF > 0.05). The significance of each variance fraction was tested with 1,000 permutations.

## Results

### Genome-wide SNP variation in Senegalese sorghum

The Senegalese sorghum accessions included in this study originated from diverse agroclimatic zones (Fig. 2-1A), agro-ecological regions (Fig. A-1A), and ethnic-linguistic groups (Fig. A-1B). Across 421 accessions, we identified 213,916 SNPs after filtering out SNPs with > 20% missing data, MAF < 0.01, and retaining only biallelic SNPs. The SNP density was determined based on non-overlapping windows of 1 Mb where SNPs were distributed across the genome with higher density in the subtelomeric regions (Fig. A-2A). The SNPs covered most of the genome with an average coverage of 1 SNP every 2 kb. The average observed and expected heterozygosity in the SSG were estimated at 0.05 and 0.23, respectively. The average pairwise nucleotide diversity ($\pi$) was 0.00054 in durra and 0.00060 in guinea accessions. The average pairwise LD ($r^2$) decreased from its initial value (~0.5) to 0.2 at 220 kb, 150 kb, and 81 kb in durra, guinea, and whole SSG, respectively (Fig. A-2B). LD decayed to background level (~0.1) at 880 kb in durra and 430 kb in guinea. The SSG had a lesser proportion of low frequency minor alleles (<5% MAF) and greater proportion of intermediate frequency minor alleles than the GDP, based on non-overlapping window size of 1 Mb (Fig. A-2C). About 60% of SNPs were rare (MAF < 0.05).

Next, we investigated the genetic variation and structure of the SSG. The two first principal components explained 3.8% and 2.5% of SNP variation (Fig. 2-1B). The accessions originated from the center formed one cluster, accessions from the south formed a second cluster, and accessions from the north formed a third cluster. The third cluster included durra accessions, caudatum accessions, a few guinea accessions from the north, and improved varieties. Neighbor joining (NJ) tree matched the PCA results and revealed that SSG durra accessions were closely

related to the durra from Ethiopia and other West African countries (Fig. A-2D). Durra and guinea accessions within the SSG were genetically differentiated from each other. The SSG also clustered somewhat with respect to ethno-linguistic groups, which are nested within geographic origins of the accessions (Fig. A-2E).

## Model-based population structure and variance partitioning

To further characterize genetic structure and gene flow among groups we used Bayesian model-based clustering. ADMIXTURE revealed a hierarchical genetic structure and high amount of gene flow among subpopulations (Supporting Information File S1). Cross validation error was minimized with K = 7 subpopulations (Fig. A-3). We investigated the spatial genetic co-ancestry in the SSG with TESS3 based on allele frequency distribution and geographic origin. Seven optimum spatial genetic clusters (K = 7) were identified (Fig. 2-2A). The TESS3 results matched the ADMIXTURE groups for different K values. Genetic differentiation among the subpopulations (including only samples with admixture rate ≥ 0.7) found at K = 7 from TESS3 results was determined using the $F_{ST}$ analysis (Fig. 2-2B). The durra accessions from the northern subpopulation (pop1) were more related to the improved varieties (pop4), based on $F_{ST}$ analysis. Both pop1 and pop4 were distinct from central and southern subpopulations (pop2, 3, 5-7), which were mostly formed by guinea accessions where guineas in the center were differentiated from guineas in the south. $F_{ST}$ of 0.185 and 0.052 were estimated between guinea and durra accessions in the SSG, and between SSG durra and GDP durra, respectively.

We used RDA to estimate the proportion on SNP variation explained by climate variation, ethno-linguistic origin, and space. Climate and ethnicity explained up to 6% and 4% of SNP variance, respectively, including variance collinear with space (P > 0.001) (Fig. 2-2C). After accounting for space, climate and ethnicity explain up to 2% of the variance, each. Climate collinear with space, the putative proportion of clinal adaptation, explained 6% of variance.

## Genome-wide patterns of nucleotide polymorphism

To identify genomic regions subject to selection, we compared genome-wide nucleotide polymorphism ($\pi$) between guinea (Soudanian) and durra (Sahelian) accessions within the SSG. Since guinea sorghums are generally more genetically diverse than durra sorghums, we used $\pi_{guinea}$ in the numerator and $\pi_{durra}$ in the denominator to identify low-diversity genomic regions in the durra genome. Nucleotide polymorphism was reduced in durra compared to guinea across most of the genome, with notably low $\pi$ on pericentromeric regions of chromosome 2, 5, 7, and

10 (Fig. 2-3A). For durra, 34 genomic regions (1 Mb windows) were identified as putative selected regions (top 5% cutoff > 1.96). A notable region of low $\pi_{durra}$ on chromosome 1 co-localized with the *Ma3* photoperiodic flowering gene. Modestly lower $\pi_{durra}$ was observed around *Stg1*, *Stg3a*, and *Stg3b*. Generally, negative values of Tajima's *D* were observed in durra, contrasting the positive values observed in guinea (Fig. 2-3B).

To better understand the timing of putative selection events, we investigated ratios of nucleotide polymorphism across three putative genetic bottlenecks: (*i*) since the divergence of durra from its common ancestor with guinea types, (*ii*) from Ethiopian durra (center of durra origin) to West African durra (Niger and Mali), (*iii*) and from West African durra to Senegalese durra (Fig. 2-3C). We also characterized nucleotide polymorphism between all Sahelian durra against worldwide guinea (Fig. A-4). The π reduction in the pericentromeric regions of chromosome 4 occurred mainly in Ethiopian durra. The π reduction on pericentromeric regions of chromosomes 5 and 10 and subtelomeric region of chromosome 6 were common to all West African durra sorghums. The π reduction in the pericentromeric region of chromosome 2 was specific to the SSG durra.

**Selective sweeps and colocalization of a priori candidate genes**

Next, we used CLR to identify candidate selective sweeps for Sahelian adaptation in durra in the SSG. CLR identified 47 candidate genomic regions (top 5% cutoff or CLR > 16.9) in durra (Fig. 2-4A). We investigated if *a priori* candidate genes (n = 64) implicated in stay-green, flowering time, or inflorescence morphology co-localized with CLR outliers. Given that the candidate genes were identified *a priori* from the literature, a liberal cutoff of 1 Mb was used to define colocalization between CLR outlier regions and candidate genes. Sixteen out of 47 CLR outliers co-localized with candidate genes (Supporting Information Data S3). The photoperiodic flowering genes *Ma3*, *GI*, *CRY1,* and *ZFL1* and inflorescence architecture candidate genes *HAM3, Sbra2,* and *vt2* colocalized with CLR outliers. The stay-green loci *Stg3a* and *Stg3b* colocalized with outlier regions on subtelomeric regions of chromosome 2. We used CLR in guinea to identify candidate selective sweeps for Soudanian adaptation. The CLR identified 28 candidate genomic regions (CLR > 10.3) in guinea (Fig. 2-4B). Eleven out of 28 CLR outliers colocalized with candidate genes (Supporting Information Data S3). The photoperiodic flowering genes *PhyA*, *Hd1, SbCN2,* and *Ma6* colocalized with outlier regions. The stay-green

locus *Stg1* colocalized with an outlier region on chromosome 3. The inflorescence morphology genes *IDS1*, *DFL2*, *Sbra3*, and *Dwarf8* colocalized with outlier regions.

## Genome-wide association studies of putative adaptive traits

To better characterize variation underlying putative adaptive traits, we mapped genotype-phenotype associations for photoperiodic flowering and inflorescence morphology. To reduce confounding effects of population structure, we also applied a regional mapping approach where durra accessions were excluded. In total, 445 and 178 significantly associated SNPs (Bonferroni $p$-value > 0.05) were identified for photoperiod sensitivity for the whole SSG and SSG without durra, respectively (Fig. 2-5A and Fig. A-5A). Colocalization between associated SNPs and candidate genes was determined based on LD decay rate to background level ($r2$ = 0.1) in durra (800 kb) and guinea (500 kb). Among the associated SNPs, 35 and 26 co-localized with photoperiodic flowering candidate genes for the whole SSG and SSG excluding durra, respectively. For panicle compactness, 48 and 124 significantly associated SNPs were found for the whole SSG and SSG excluding durra, respectively (Fig. 2-5B and Fig. A-5B). Among the associated SNPs, 13 SNPs co-localized with *a priori* candidate genes for inflorescence morphology.

Photoperiod sensitivity-associated SNPs were found near floral regulators *Ma3*, *Ma5*, *Ma6*, *MADS14*, *GI*, *HD6*, *zfl1/2*, *Ehd2*, *SbCN12*, and *SbCN15* (Supporting Information Data S4). Most of these associations were observed whether or not durra were included. The association near *Ehd2* was only observed when durra accessions were excluded, while associations near *Ma6* and *HD6* were only observed when durra accessions were included. Eighteen of the highly significant ($p$-value > $10^{-10}$) associations were not near any *a priori* candidate genes. For panicle compactness, significantly associated SNPs co-localized with *SP1*, *CRCK3*, *TCP24*, *DFL2*, *Sb-ra2*, *vt2*, and *rel2*. The SNP S1_55302939 (within the *SP1* gene) was significant in both GWAS approaches, while S1_55305415 (1 kb away from *SP1*) was only significant when using the whole SSG panel. Two of the highly significant ($p$-value > $10^{-10}$) associations were not near *a priori* candidate genes.

## Environment-SNP associations

We performed genome-environment associations to identify SNPs associated with climate variables (Supporting Information Data S4). Based on the GLM, genome-environment associations identified 560 SNPs significantly associated (Bonferroni-adjusted $p$-value > 0.05)

with environment variables including precipitation of the driest quarter (Fig. 2-6A), mean temperature of the warmest quarter (Fig. A-6A), and precipitation of the wettest quarter (Fig. A-6B). Associations for longitude variable were based on the MLM (Fig. A-6C) because GLM identified many associated SNPs. MLMM identified 16 significantly associated SNPs, including one overlapping SNP (S7_59683060) with the GLM, and 15 additional SNPs that were not identified by GLM or MLM (Table A-1, Fig. 2-6A, and Fig. A-6). Associated SNPs for precipitation of the driest quarter, such as S2_60708848 and S6_691400, identified by the MLMM, colocalized with the *Stg3a* locus and *Ma6* gene, respectively. The stay-green candidate loci (*Stg1–4*) co-localized with SNPs associated with mean temperature of the driest and warmest quarters, precipitation of the driest, warmest and wettest quarters, and longitude (Supporting Information Data S4). The SNP S1_7584419 identified by MLMM as associated with mean temperature of the warmest quarter colocalized with *Ma5* and *MADS14*, but at greater distance (> 800 kb).

To determine the pairwise LD between the two SNPs co-localizing with *Stg3a* and *Ma6* and variation within these loci, we generated the LD heatmap of the 50 kb region surrounding each SNP (Fig. 2-6B, C). Nearly complete LD ($r^2 > 0.9$) was found between S2_60708848 and other SNPs in the *Stg3a* locus. The SNP S6_691400 was in LD with two SNPs in *Ma6*. The genotypes carrying the minor allele at S2_60708848 were distributed in the southern subhumid environments (Fig. 2-6D). By contrast, genotypes carrying the minor allele at S6_691400 were distributed in the northern and dry environments (Fig. 2-6E). The minor alleles at S3_67831630 (co-localized with *Stg1/SbPIN4*) and S3_57321183 (co-localized with *Stg2/SbPIN2*) were mostly found in durra landraces and few guinea landraces distributed in the dry areas of Senegal (Fig. A-7A, B).

## Discussion

Genomic analysis of crop landraces can help determine the basis of local adaptation (Lasky et al., 2015; Li et al., 2017; Lin et al., 2014; Swarts et al., 2017). The aims of this study were to characterize factors shaping the genomic variation of Senegalese sorghum landraces, map genomic regions shaped by agroclimatic adaptation, and identify genes that could play a role in local adaptation.

## Factors shaping genomic variation of sorghum landraces

Population structure in the Senegalese landraces followed the north-south precipitation gradient. These regional-scale patterns are in line with global patterns, where population structure is associated with precipitation-based agroclimatic zones (Lasky et al., 2015). Within Senegalese sorghums, guinea and durra clustered distinctly, consistent with global patterns of genetic differentiation (Morris et al., 2013; Sagnard et al., 2011). The relatively high proportion of variation explained by climate collinear with space suggests a role of clinal adaptation shaping variation, similar to recent findings in Nigerian and global sorghum germplasm (Lasky et al., 2015; Olatoye et al., 2018). However, two guinea groups, from the center and south, clustered distinctly (Fig. 2-1B and Fig. A-2D), suggesting possibly a specific genomic adaptation to the Soudano-Sahelian and the Soudanian agro-climatic, respectively.

The average pairwise nucleotide diversity, observed heterozygosity (data not shown), and the spatial and hierarchical genetic structure observed within guinea group (Fig. 2-2A; Supporting Information File S1) is consistent with guinea being the most genetically diverse sorghum type (Deu et al., 1994; Folkertsma et al., 2005; Morris et al., 2013). Although the number of inferred subpopulations may not always correspond to the number of biological genetic groups (François and Durand, 2010; Meirmans, 2015), the spatial genetic co-ancestry structure analysis suggests the presence of untapped genetic diversity in the subpopulations in eastern Senegal (Fig. 2-2A). The high estimated admixture coefficients among putative guinea subpopulations (Supporting Information File S1) could be due to gene flow among subpopulations or an effect of limited isolation-by-distance. The limited isolation-by-distance may occur because the geographic origin of the accessions is not broad (e.g. Senegal is not large) and there is any major geographic barrier that may create isolated subpopulations. There was little evidence of admixture between guinea and durra types, consistent with phenotype studies that rarely identify guinea-durra intermediates (Harlan and De Wet, 1972). Evidence of gene flow was mostly from guinea to durra (e.g. red subgroup at K = 7, Supporting Information File S1) and rarely from durra to guinea. The lower abundance of durra in this region may explain the limited gene flow between guinea and durra sorghums.

Ethnicity of farmers has shaped genetic structure in several staple crops including maize and pearl millet. Distribution and diffusion of ethnic groups in Senegal including the Wolof, Serer, Diola, and Fulani (Toucouleur, Peul, Peul Foulbe, and Peul Firdou) could have affected

gene flow among landraces. Indeed, the ethno-linguistic origin of the accessions contributed to the genetic variance of the Senegalese sorghum (Fig. 2-2C). Seed exchange among farmers of the same ethnic group may have contributed in shaping this genetic structure (Barnaud, Trigueros, McKey, & Joly, 2008; Orozco-Ramírez et al., 2016; Pressoir & Berthaud, 2004)(Barnaud et al., 2008; Orozco-Ramírez et al., 2016; Pressoir and Berthaud, 2004). Co-diffusion of sorghum with human migration has been demonstrated at Africa-wide scale (Westengen et al., 2014) and at a regional scale in Kenya (Labeyrie et al., 2016). Durra sorghum in Senegal are grown mainly by the Fulani ethnic group, so the clustering of Senegalese durra with Ethiopian durra (Fig. A-2D) and low $F_{ST}$ (0.052) suggest that durra sorghums moved with Fulani people from northeast Africa (Scheinfeldt et al., 2010).

## Genetic basis of Sahelian and Soudanian adaptation

Nucleotide polymorphism patterns can provide insight into loci underlying adaptation (Vitti et al., 2013). The reduction of nucleotide polymorphism observed throughout the durra genome (Fig. 2-3A) could be resulted from the bottlenecks during its diffusion along the Sahelian zone. Because Ethiopia is known as the center of origin of durra, we investigated whether the reduced polymorphism in durra was common to all African durra or specific to the Senegalese durra. The results suggest selective sweeps across durra genomes as durra populations diffused along the Sahel (Fig. 2-3C and Fig. A-4). Interestingly, putative selective sweeps on pericentromeric regions of chromosome 2 were specific to Senegalese durra. By contrast to durra, there was little reduction of nucleotide polymorphism in the guinea genome and predominantly positive values of Tajima's *D* test (Fig. 2-3A, B), reflecting population structure or possible balancing selection (Vitti et al., 2013). Simulations with demographic models could be used for more robust genome scans. Unfortunately, the underlying population parameters (e.g. effective population size, migration rates) are poorly described in sorghum.

Photoperiodic flowering is a key factor underlying adaptation in tropical crops (Kloosterman et al., 2013). The co-localization of photoperiodic flowering candidate genes with putative selective sweeps and phenotypic and environment associations (Fig. 2-4, 5A; Table A-1; Supporting Information Data S4) are consistent with a role of conserved cereal flowering pathways in sorghum climate adaptation. The rare allele at the SNP near *Ma6/Ghd7* (6 kb away) was present in durra genotypes distributed in the drier areas of the Sahelian zone characterized by short growing seasons and low rainfall (< 400 mm per year) (Fig. 2-6E). This rare allele may

be associated with early maturity and thus suggesting a role in drought escape such that plants can rapidly cover their maturity cycle and produce seeds before the end of growing season.

Other photoperiod flowering regulators identified in U.S. sorghum, *PhyC* (*Ma5*), *PhyB* (*Ma3*), *PhyA*, and *Ma1* (*SbPRR37*), co-localized with phenotype-associated SNPs (Supporting Information Data S4) and putative selective sweeps in the Senegalese sorghum (Childs et al., 1997; Rooney and Aydin, 1999). Signatures of selection near *Ma3* in durra ($\pi_{durra}$ and CLR; Fig. 2-3A, 4A) are consistent with signatures of positive selection in *Ma3* observed in global sorghum (Wang et al., 2015). The florigens *SbCN12* and *SbCN15* (ortholog of rice florigen *Hd3a*) found near photoperiod-associated SNPs are photoperiod-regulated activators of floral induction in U.S. sorghum (Murphy et al., 2011). Putative photoperiodic flowering regulators *SbCRY1* and *SbGI*, co-localizing with selective sweeps in durra and photoperiod-associated SNPs, were previously associated with photoperiodic flowering in regional West-Central African germplasm (Bhosale et al., 2012). Several of the above genes were associated with flowering time adaptation in maize landraces, including *PhyB*, *PhyC*, *PRR37*, and *ZFL1/2* (Romero Navarro et al., 2017).

Panicle compactness in sorghum is a function of the number and length of inflorescence branches and the number of aborted spikelets (Brown et al., 2006). Several candidate genes from a previous GWAS of inflorescence branch length in global sorghum (Morris et al., 2013) co-localized with GWAS signals for panicle compactness and/or CLR outliers in the current study (*SP1*, *CRCK3/THE1*, *TCP24*, and *DFL2*). The minor alleles in/near *SP1* (S1_55302939, S1_55305415) observed in durra accessions and some guinea accessions (Fig. A-7C–D) suggests a rare variant in *SP1* could contribute to shorter inflorescence branches in some Senegalese sorghum.

The colocalization of selective sweeps and genome-environment associations (Supporting Information Data S4) with stay-green drought tolerance loci (Borrell et al., 2014) suggests a broader role for stay-green loci in Sahelian adaptation. A selective sweep and associated SNPs colocalized with the stay-green locus *Stg1*/*SbPIN4* in guinea sorghums, suggesting that this region may confer adaptation of some guinea accessions to the dry areas of Senegal. The rare allele of SNP S3_57321183, which co-localized with *SbPIN2,* was found in a few guinea sorghums (Figure A-7B). One possibility is that severe droughts starting in the 1970s (Gautier et al., 2016; Mbow et al., 2008) have favored the introgression of stay-green drought tolerance alleles into some guinea landraces. Genome scans comparing older landrace collections with

recent collections may shed more light on whether more recent selection (e.g. 1970s-2000s) has occurred, as demonstrated in Sahelian pearl millet (Vigouroux et al., 2011).

## Prospects for genomic dissection and improvement of climate adaptation

Improving adaptation of staple crops to the Sahelian and Soudanian zones is critical for smallholder farmers and a major challenge for African plant breeders. Despite advances in genotyping platforms, genomic tools for crop adaptation in sub-Saharan countries remain lacking. This study generated substantial genomic resources (213,916 SNPs among which 145,235 SNPs have MAF > 0.05) representing high quality-markers useful for the genomic dissection of adaptive and complex traits. High rates of SNPs with low frequency minor alleles (about 60 % of the data had MAF < 0.05) were detected. One possible explanation may be related to the fact that these accessions are mostly landraces grown in their center of origin; thus high number of rare polymorphisms might be segregating at intermediate frequency in the germplasm. In the USDA-NPGS Ethiopian sorghum collection, similar patterns of MAF were found where 60% of detected SNPs had MAF < 0.05 (Cuevas et al., 2017). Overall, the Senegalese sorghum landraces represent a useful genetic resource, harboring useful variation for maturity and inflorescence morphology, as well as resistant sources to grain mold and anthracnose (Cuevas et al., 2018).

The moderate decay of LD observed within the germplasm (Fig. A-2B) is consistent with the predominance of inbreeding in sorghum (Hamblin et al., 2005). Studies in sorghum have found a comparable LD pattern, decaying to its background level at ~150 kb (Mace et al., 2013; Morris et al., 2013). The population structure of the Senegalese sorghum landraces would be expected to increase spurious association and reduce the power of GWAS (Brachi et al., 2011). Indeed, the number of associations for photoperiod sensitivity was reduced when applying the regional mapping approach excluding durra accessions, presumably due to fewer spurious associations. Future studies with West African multi-parent mapping populations could break-up confounding LD and improve power to detect climate-adaptive loci (McMullen et al., 2009; Bouchet et al., 2017).

The stay-green loci may be useful to improve for drought adaptation in the Sahel via marker-assisted selection. Circadian clock-related genes influence crop yield under abiotic stress (Bendix et al., 2015) and photoperiodic flowering loci identified may contribute to early maturity and drought escape in the Sahel. Taken together, our findings suggest a complex oligogenic basis

of adaptation to Sahelian versus Soudanian climate, underpinned by variation in conserved floral regulatory pathways and variation in other pathways that are more poorly understood. Whole-genome resequencing of African crop diversity for GWAS and genome scans could facilitate identification of causal variants in the molecular pathways that underlie climate adaptation.

## Acknowledgements

This study is made possible by the support of the American People provided to the Feed the Future Innovation Lab for Collaborative Research on Sorghum and Millet through the United States Agency for International Development (USAID) under Cooperative Agreement No. AID-OAA-A-13-00047. The contents are the sole responsibility of the authors and do not necessarily reflect the views of USAID or the United States Government. The study was conducted using resources at the Integrated Genomics Facility and Beocat High Performance Computing Cluster at Kansas State University. This study is contribution number *[in process]* of the Kansas Agricultural Experiment Station.

## Data Accessibility

The raw sequencing data generated in this study are available in NCBI under the accession number SRP132525. The SNP data set is available at Dryad Data Repository under accession https://doi.org/10.5061/dryad.32f5395.

## Supporting Information Available

Additional supporting information may be found online in the Supporting Information section at the end of the article. *Ecol Evol*. 2019;00:1–14.
https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.5187

Supporting Information Data S1 contains information about the sorghum accessions.

Supporting Information Data S2 contains information about the candidate genes, gene functions, and references.

Supporting Information Data S3 contains CLR outliers in durra and guinea sorghum accessions.

Supporting Information Data S4 contains significantly associated SNPs with phenotypes or environment variables.

Supporting Information File S1 contains ADMIXTURE population structure results at K = 3–7.

## References


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

Barnaud, A., Trigueros, G., McKey, D., Joly, H.I., 2008. High outcrossing rates in fields with mixed sorghum landraces: how are landraces maintained? Heredity 101, 445. https://doi.org/10.1038/hdy.2008.77

Bendix, C., Marshall, C.M., Harmon, F.G., 2015. Circadian Clock Genes Universally Control Key Agricultural Traits. Molecular Plant 8, 1135–1152. https://doi.org/10.1016/j.molp.2015.03.003

Bhosale, S.U., Stich, B., Rattunde, H.F.W., Weltzien, E., Haussmann, B.I., Hash, C.T., Ramu, P., Cuevas, H.E., Paterson, A.H., Melchinger, A.E., Parzies, H.K., 2012. Association analysis of photoperiodic flowering time genes in west and central African sorghum [Sorghum bicolor (L.) Moench]. BMC Plant Biology 12, 32. https://doi.org/10.1186/1471-2229-12-32

Blum, A., 2014. Genomics for drought resistance – getting down to earth. Functional Plant Biology 41. https://doi.org/10.1071/FP14018

Borrell, A.K., van Oosterom, E.J., Mullet, J.E., George-Jaeggli, B., Jordan, D.R., Klein, P.E., Hammer, G.L., 2014. Stay-green alleles individually enhance grain yield in sorghum under drought by modifying canopy development and water uptake patterns. New Phytol 203, 817–830. https://doi.org/10.1111/nph.12869

Bouchet, S., Olatoye, M.O., Marla, S.R., Perumal, R., Tesso, T., Yu, J., Tuinstra, M., Morris, G.P., 2017. Increased Power To Dissect Adaptive Traits in Global Sorghum Diversity Using a Nested Association Mapping Population. Genetics 206, 573–585. https://doi.org/10.1534/genetics.116.198499

Bouchet, S., Pot, D., Deu, M., Rami, J.-F., Billot, C., Perrier, X., Rivallan, R., Gardes, L., Xia, L., Wenzl, P., Kilian, A., Glaszmann, J.-C., 2012. Genetic Structure, Linkage Disequilibrium and Signature of Selection in Sorghum: Lessons from Physically Anchored DArT Markers. PLOS ONE 7, e33470. https://doi.org/10.1371/journal.pone.0033470

Brachi, B., Morris, G.P., Borevitz, J.O., 2011. Genome-wide association studies in plants: the missing heritability is in the field. Genome Biology 12, 232. https://doi.org/10.1186/gb-2011-12-10-232

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y., Buckler, E.S., 2007. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23, 2633–2635. https://doi.org/10.1093/bioinformatics/btm308

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

Caicedo, A.L., Williamson, S.H., Hernandez, R.D., Boyko, A., Fledel-Alon, A., York, T.L., Polato, N.R., Olsen, K.M., Nielsen, R., McCouch, S.R., Bustamante, C.D., Purugganan, M.D., 2007. Genome-Wide Patterns of Nucleotide Polymorphism in Domesticated Rice. PLOS Genetics 3, e163. https://doi.org/10.1371/journal.pgen.0030163

Camus-Kulandaivelu, L., Veyrieras, J.-B., Madur, D., Combes, V., Fourmann, M., Barraud, S., Dubreuil, P., Gouesnard, B., Manicacci, D., Charcosset, A., 2006. Maize Adaptation to Temperate Climate: Relationship Between Population Structure and Polymorphism in the Dwarf8 Gene. Genetics 172, 2449–2463. https://doi.org/10.1534/genetics.105.048603

Cavanagh, C., Morell, M., Mackay, I., Powell, W., 2008. From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. Current Opinion in Plant Biology, Genome studies and Molecular Genetics, edited by Juliette de Meaux and Maarten Koornneef / Plant Biotechnology, edited by Andy Greenland and Jan Leach 11, 215–221. https://doi.org/10.1016/j.pbi.2008.01.002

Caye, K., Deist, T.M., Martins, H., Michel, O., François, O., 2016. TESS3: fast inference of spatial population structure and genome scans for selection. Mol Ecol Resour 16, 540–548. https://doi.org/10.1111/1755-0998.12471

Childs, K.L., Miller, F.R., Cordonnier-Pratt, M.M., Pratt, L.H., Morgan, P.W., Mullet, J.E., 1997. The sorghum photoperiod sensitivity gene, Ma3, encodes a phytochrome B. Plant Physiol 113, 611–619.

Clément, J.-C., Houdiard, P., 1977. Prospection des Mils pénicillaires et Sorghos en Afrique de l'Ouest: Campagne 1976: Nigeria-Sénégal.

Cuevas, H.E., Prom, L.K., Rosa-Valentin, G., 2018. Population structure of the NPGS Senegalese sorghum collection and its evaluation to identify new disease resistant genes. PLOS ONE 13, e0191877. https://doi.org/10.1371/journal.pone.0191877

Cuevas, H.E., Rosa-Valentin, G., Hayes, C.M., Rooney, W.L., Hoffmann, L., 2017. Genomic characterization of a core set of the USDA-NPGS Ethiopian sorghum germplasm collection: implications for germplasm conservation, evaluation, and utilization in crop improvement. BMC Genomics 18, 108. https://doi.org/10.1186/s12864-016-3475-7

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R., Group, 1000 Genomes Project Analysis, 2011. The variant call format and VCFtools. Bioinformatics 27, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

de Meeûs, T., Goudet, J., 2007. A step-by-step tutorial to use HierFstat to analyse populations hierarchically structured at multiple levels. Infection, Genetics and Evolution 7, 731–735. https://doi.org/10.1016/j.meegid.2007.07.005

Deu, M., Gonzalez-de-Leon, D., Glaszmann, J.-C., Degremont, I., Chantereau, J., Lanaud, C., Hamon, P., 1994. RFLP diversity in cultivated sorghum in relation to racial differentiation. Theoret. Appl. Genetics 88, 838–844. https://doi.org/10.1007/BF01253994

Doggett, H., 1988. Sorghum. Longman Scientific & Technical.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., Mitchell, S.E., 2011. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE 6, e19379. https://doi.org/10.1371/journal.pone.0019379

Fang, L., Wang, Q., Hu, Y., Jia, Y., Chen, J., Liu, B., Zhang, Z., Guan, X., Chen, S., Zhou, B., Mei, G., Sun, J., Pan, Z., He, S., Xiao, S., Shi, W., Gong, W., Liu, J., Ma, J., Cai, C., Zhu, X., Guo, W., Du, X., Zhang, T., 2017. Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits. Nat Genet 49, 1089–1098. https://doi.org/10.1038/ng.3887

Folkertsma, R.T., Rattunde, H.F.W., Chandra, S., Raju, G.S., Hash, C.T., 2005. The pattern of genetic diversity of Guinea-race Sorghum bicolor (L.) Moench landraces as revealed with SSR markers. Theor Appl Genet 111, 399–409. https://doi.org/10.1007/s00122-005-1949-0

Fournier-Level, A., Korte, A., Cooper, M.D., Nordborg, M., Schmitt, J., Wilczek, A.M., 2011. A Map of Local Adaptation in Arabidopsis thaliana. Science 334, 86–89. https://doi.org/10.1126/science.1209271

François, O., Durand, E., 2010. Spatially explicit Bayesian clustering models in population genetics. Molecular Ecology Resources 10, 773–784. https://doi.org/10.1111/j.1755-0998.2010.02868.x

Franks, S.J., Sim, S., Weis, A.E., 2007. Rapid evolution of flowering time by an annual plant in response to a climate fluctuation. PNAS 104, 1278–1282. https://doi.org/10.1073/pnas.0608379104

Gautier, D., Denis, D., Locatelli, B., 2016. Impacts of drought and responses of rural populations in West Africa: a systematic review. Wiley Interdisciplinary Reviews: Climate Change 7, 666–681. https://doi.org/10.1002/wcc.411

Glaubitz, J.C., Casstevens, T.M., Lu, F., Harriman, J., Elshire, R.J., Sun, Q., Buckler, E.S., 2014. TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. PLOS ONE 9, e90346. https://doi.org/10.1371/journal.pone.0090346

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., Rokhsar, D.S., 2012. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res 40, D1178–D1186. https://doi.org/10.1093/nar/gkr944

Hamblin, M.T., Fernandez, M.G.S., Casa, A.M., Mitchell, S.E., Paterson, A.H., Kresovich, S., 2005. Equilibrium Processes Cannot Explain High Levels of Short- and Medium-Range Linkage Disequilibrium in the Domesticated Grass Sorghum bicolor. Genetics 171, 1247–1256. https://doi.org/10.1534/genetics.105.041566

Harlan, J.R., De Wet, J.J.M., 1972. A Simplified Classification of Cultivated Sorghum1. Crop Science 12, 172–176. https://doi.org/10.2135/cropsci1972.0011183X001200020005x

Harris, K., Subudhi, P.K., Borrell, A., Jordan, D., Rosenow, D., Nguyen, H., Klein, P., Klein, R., Mullet, J., 2007. Sorghum stay-green QTL individually reduce post-flowering drought-induced leaf senescence. J Exp Bot 58, 327–338. https://doi.org/10.1093/jxb/erl225

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25, 1965–1978. https://doi.org/10.1002/joc.1276

Hill, W.G., Weir, B.S., 1988. Variances and covariances of squared linkage disequilibria in finite populations. Theoretical Population Biology 33, 54–78. https://doi.org/10.1016/0040-5809(88)90004-4

Kebede, H., Subudhi, P.K., Rosenow, D.T., Nguyen, H.T., 2001. Quantitative trait loci influencing drought tolerance in grain sorghum (Sorghum bicolor L. Moench). Theor Appl Genet 103, 266–276. https://doi.org/10.1007/s001220100541

Kenney, A.M., McKay, J.K., Richards, J.H., Juenger, T.E., 2014. Direct and indirect selection on flowering time, water-use efficiency (WUE, δ13C), and WUE plasticity to drought in Arabidopsis thaliana. Ecol Evol 4, 4505–4521. https://doi.org/10.1002/ece3.1270

Kloosterman, B., Abelenda, J.A., Gomez, M. del M.C., Oortwijn, M., de Boer, J.M., Kowitwanich, K., Horvath, B.M., van Eck, H.J., Smaczniak, C., Prat, S., Visser, R.G.F., Bachem, C.W.B., 2013. Naturally occurring allele diversity allows potato cultivation in northern latitudes. Nature 495, 246–250. https://doi.org/10.1038/nature11912

Labeyrie, V., Thomas, M., Muthamia, Z.K., Leclerc, C., 2016. Seed exchange networks, ethnicity, and sorghum diversity. Proc Natl Acad Sci U S A 113, 98–103. https://doi.org/10.1073/pnas.1513238112

Lasky, J.R., Des Marais, D.L., McKay, J.K., Richards, J.H., Juenger, T.E., Keitt, T.H., 2012. Characterizing genomic variation of Arabidopsis thaliana: the roles of geography and climate. Molecular Ecology 21, 5512–5529. https://doi.org/10.1111/j.1365-294X.2012.05709.x

Lasky, J.R., Upadhyaya, H.D., Ramu, P., Deshpande, S., Hash, C.T., Bonnette, J., Juenger, T.E., Hyma, K., Acharya, C., Mitchell, S.E., Buckler, E.S., Brenton, Z., Kresovich, S., Morris, G.P., 2015. Genome-environment associations in sorghum landraces predict adaptive traits. Science Advances 1, e1400218. https://doi.org/10.1126/sciadv.1400218

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, L.-F., Li, Y.-L., Jia, Y., Caicedo, A.L., Olsen, K.M., 2017. Signatures of adaptation in the weedy rice genome. Nat Genet 49, 811–814. https://doi.org/10.1038/ng.3825

Lin, T., Zhu, G., Zhang, J., Xu, X., Yu, Q., Zheng, Z., Zhang, Z., Lun, Y., Li, S., Wang, X., Huang, Z., Li, Junming, Zhang, C., Wang, T., Zhang, Yuyang, Wang, A., Zhang, Yancong, Lin, K., Li, C., Xiong, G., Xue, Y., Mazzucato, A., Causse, M., Fei, Z., Giovannoni, J.J., Chetelat, R.T., Zamir, D., Städler, T., Li, Jingfu, Ye, Z., Du, Y., Huang, S., 2014. Genomic analyses provide insights into the history of tomato breeding. Nat Genet 46, 1220–1226. https://doi.org/10.1038/ng.3117

Lipka, A.E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P.J., Gore, M.A., Buckler, E.S., Zhang, Z., 2012. GAPIT: genome association and prediction integrated tool. Bioinformatics 28, 2397–2399. https://doi.org/10.1093/bioinformatics/bts444

Mace, E.S., Tai, S., Gilding, E.K., Li, Y., Prentis, P.J., Bian, L., Campbell, B.C., Hu, W., Innes, D.J., Han, X., Cruickshank, A., Dai, C., Frère, C., Zhang, H., Hunt, C.H., Wang, X., Shatte, T., Wang, M., Su, Z., Li, J., Lin, X., Godwin, I.D., Jordan, D.R., Wang, J., 2013. Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. Nat Commun 4, 2320. https://doi.org/10.1038/ncomms3320

Mbow, C., Mertz, O., Diouf, A., Rasmussen, K., Reenberg, A., 2008. The history of environmental change and adaptation in eastern Saloum–Senegal—Driving forces and perceptions. Global and Planetary Change, Climate Change and Desertification 64, 210–221. https://doi.org/10.1016/j.gloplacha.2008.09.008

McCormick, R.F., Truong, S.K., Sreedasyam, A., Jenkins, J., Shu, S., Sims, D., Kennedy, M., Amirebrahimi, M., Weers, B., McKinley, B., Mattison, A., Morishige, D., Grimwood, J., Schmutz, J., Mullet, J., 2017. The Sorghum bicolor reference genome: improved assembly and annotations, a transcriptome atlas, and signatures of genome organization. bioRxiv 110593. https://doi.org/10.1101/110593

McCormick, R.F., Truong, S.K., Sreedasyam, A., Jenkins, J., Shu, S., Sims, D., Kennedy, M., Amirebrahimi, M., Weers, B.D., McKinley, B., Mattison, A., Morishige, D.T., Grimwood, J., Schmutz, J., Mullet, J.E., 2018. The Sorghum bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. The Plant Journal 93, 338–354. https://doi.org/10.1111/tpj.13781

McMullen, M.D., Kresovich, S., Villeda, H.S., Bradbury, P., Li, H., Sun, Q., Flint-Garcia, S., Thornsberry, J., Acharya, C., Bottoms, C., Brown, P., Browne, C., Eller, M., Guill, K., Harjes, C., Kroon, D., Lepak, N., Mitchell, S.E., Peterson, B., Pressoir, G., Romero, S., Rosas, M.O., Salvo, S., Yates, H., Hanson, M., Jones, E., Smith, S., Glaubitz, J.C., Goodman, M., Ware, D., Holland, J.B., Buckler, E.S., 2009. Genetic Properties of the Maize Nested Association Mapping Population. Science 325, 737–740. https://doi.org/10.1126/science.1174320

Meirmans, P.G., 2015. Seven common mistakes in population genetics and how to avoid them. Mol Ecol 24, 3223–3231. https://doi.org/10.1111/mec.13243

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

Morton, J.F., 2007. The impact of climate change on smallholder and subsistence agriculture. PNAS 104, 19680–19685. https://doi.org/10.1073/pnas.0701855104

Mullet, J.E., Rooney, W.L., Klein, P.E., Morishige, D., Murphy, R., Brady, J.A., 2010. Discovery and utilization of sorghum genes (ma5/ma6). US20100024065 A1.

Murphy, R.L., Klein, R.R., Morishige, D.T., Brady, J.A., Rooney, W.L., Miller, F.R., Dugas, D.V., Klein, P.E., Mullet, J.E., 2011. Coincident light and clock regulation of pseudoresponse regulator protein 37 (PRR37) controls photoperiodic flowering in sorghum. PNAS 108, 16469–16474. https://doi.org/10.1073/pnas.1106212108

Naino Jika, A.K., Dussert, Y., Raimond, C., Garine, E., Luxereau, A., Takvorian, N., Djermakoye, R.S., Adam, T., Robert, T., 2017. Unexpected pattern of pearl millet genetic diversity among ethno-linguistic groups in the Lake Chad Basin. Heredity 118, 491–502. https://doi.org/10.1038/hdy.2016.128

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M.J., Clark, A.G., Bustamante, C., 2005. Genomic scans for selective sweeps using SNP data. Genome Res 15, 1566–1575. https://doi.org/10.1101/gr.4252305

Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs, E., Wagner, H., 2017. vegan: Community Ecology Package.

Olatoye, M.O., Hu, Z., Maina, F., Morris, G.P., 2018. Genomic Signatures of Adaptation to a Precipitation Gradient in Nigerian Sorghum. G3: Genes, Genomes, Genetics g3.200551.2018. https://doi.org/10.1534/g3.118.200551

Olsen, K.M., Caicedo, A.L., Polato, N., McClung, A., McCouch, S., Purugganan, M.D., 2006. Selection Under Domestication: Evidence for a Sweep in the Rice Waxy Genomic Region. Genetics 173, 975–983. https://doi.org/10.1534/genetics.106.056473

Orozco-Ramírez, Q., Ross-Ibarra, J., Santacruz-Varela, A., Brush, S., 2016. Maize diversity associated with social origin and environmental variation in Southern Mexico. Heredity. https://doi.org/10.1038/hdy.2016.10

Paradis, E., Claude, J., Strimmer, K., 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20, 289–290.

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., Grigoriev, I.V., Lyons, E., Maher, C.A., Martis, M., Narechania, A., Otillar, R.P., Penning, B.W., Salamov, A.A., Wang, Y., Zhang, L., Carpita, N.C., Freeling, M., Gingle, A.R., Hash, C.T., Keller, B., Klein, P., Kresovich, S., McCann, M.C., Ming, R., Peterson, D.G., Mehboob-ur-Rahman, Ware, D., Westhoff, P., Mayer, K.F.X., Messing, J., Rokhsar, D.S., 2009. The Sorghum bicolor genome and the diversification of grasses. Nature 457, 551–556. https://doi.org/10.1038/nature07723

Pavlidis, P., Živković, D., Stamatakis, A., Alachiotis, N., 2013. SweeD: Likelihood-Based Detection of Selective Sweeps in Thousands of Genomes. Mol Biol Evol 30, 2224–2234. https://doi.org/10.1093/molbev/mst112

Pressoir, G., Berthaud, J., 2004. Patterns of population structure in maize landraces from the Central Valleys of Oaxaca in Mexico. Heredity 92, 88. https://doi.org/10.1038/sj.hdy.6800387

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C., 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics 81, 559–575. https://doi.org/10.1086/519795

R Core Team, R.C., 2016. A language and environment for statistical computing. R Foundation for statistical computing, 2015; Vienna, Austria.

Rellstab, C., Gugerli, F., Eckert, A.J., Hancock, A.M., Holderegger, R., 2015. A practical guide to environmental association analysis in landscape genomics. Mol Ecol 24, 4348–4370. https://doi.org/10.1111/mec.13322

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

Romero Navarro, J.A., Willcox, M., Burgueño, J., Romay, C., Swarts, K., Trachsel, S., Preciado, E., Terron, A., Delgado, H.V., Vidal, V., Ortega, A., Banda, A.E., Montiel, N.O.G., Ortiz-Monasterio, I., Vicente, F.S., Espinoza, A.G., Atlin, G., Wenzl, P., Hearne, S., Buckler, E.S., 2017. A study of allelic diversity underlying flowering-time adaptation in maize landraces. Nat Genet 49, 476–480. https://doi.org/10.1038/ng.3784

Rooney, W.L., Aydin, S., 1999. Genetic Control of a Photoperiod-Sensitive Response in Sorghum bicolor (L.) Moench. Crop Science 39, 397. https://doi.org/10.2135/cropsci1999.0011183X0039000200016x

Sagnard, F., Deu, M., Dembélé, D., Leblois, R., Touré, L., Diakité, M., Calatayud, C., Vaksmann, M., Bouchet, S., Mallé, Y., Togola, S., Traoré, P.C.S., 2011. Genetic diversity, structure, gene flow and evolutionary relationships within the Sorghum bicolor

wild-weedy-crop complex in a western African region. Theor. Appl. Genet. 123, 1231–1246. https://doi.org/10.1007/s00122-011-1662-0

Sanon, M., Hoogenboom, G., Traoré, S.B., Sarr, B., Garcia, A.G. y, Somé, L., Roncoli, C., 2014. Photoperiod sensitivity of local millet and sorghum varieties in West Africa. NJAS - Wageningen Journal of Life Sciences 68, 29–39. https://doi.org/10.1016/j.njas.2013.11.004

Scheinfeldt, L.B., Soi, S., Tishkoff, S.A., 2010. Working toward a synthesis of archaeological, linguistic, and genetic data for inferring African population history. PNAS 107, 8931–8938. https://doi.org/10.1073/pnas.1002563107

Segura, V., Vilhjálmsson, B.J., Platt, A., Korte, A., Seren, Ü., Long, Q., Nordborg, M., 2012. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet 44, 825–830. https://doi.org/10.1038/ng.2314

Shin, J.-H., Blay, S., McNeney, B., Graham, J., 2006. LDheatmap: An R Function for Graphical Display of Pairwise Linkage Disequilibria Between Single Nucleotide Polymorphisms | Shin | Journal of Statistical Software. J Stat Soft 16. https://doi.org/10.18637/jss.v016.c03

Siepielski, A.M., Morrissey, M.B., Buoro, M., Carlson, S.M., Caruso, C.M., Clegg, S.M., Coulson, T., DiBattista, J., Gotanda, K.M., Francis, C.D., Hereford, J., Kingsolver, J.G., Augustine, K.E., Kruuk, L.E.B., Martin, R.A., Sheldon, B.C., Sletvold, N., Svensson, E.I., Wade, M.J., MacColl, A.D.C., 2017. Precipitation drives global variation in natural selection. Science 355, 959–962. https://doi.org/10.1126/science.aag2773

Slatkin, M., 2008. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. Nat Rev Genet 9, 477–485. https://doi.org/10.1038/nrg2361

Swarts, K., Gutaker, R.M., Benz, B., Blake, M., Bukowski, R., Holland, J., Kruse-Peeples, M., Lepak, N., Prim, L., Romay, M.C., Ross-Ibarra, J., Sanchez-Gonzalez, J. de J., Schmidt, C., Schuenemann, V.J., Krause, J., Matson, R.G., Weigel, D., Buckler, E.S., Burbano, H.A., 2017. Genomic estimation of complex traits reveals ancient maize adaptation to temperate North America. Science 357, 512–515. https://doi.org/10.1126/science.aam9425

Tuinstra, M.R., Grote, E.M., Goldsbrough, P.B., Ejeta, G., 1997. Genetic analysis of post-flowering drought tolerance and components of grain development in Sorghum bicolor (L.) Moench. Molecular Breeding 3, 439–448. https://doi.org/10.1023/A:1009673126345

Vigouroux, Y., Mariac, C., Mita, S.D., Pham, J.-L., Gérard, B., Kapran, I., Sagnard, F., Deu, M., Chantereau, J., Ali, A., Ndjeunga, J., Luong, V., Thuillet, A.-C., Saïdou, A.-A., Bezançon, G., 2011. Selection for Earlier Flowering Crop Associated with Climatic Variations in the Sahel. PLOS ONE 6, e19563. https://doi.org/10.1371/journal.pone.0019563

Vitti, J.J., Grossman, S.R., Sabeti, P.C., 2013. Detecting Natural Selection in Genomic Data. Annual Review of Genetics 47, 97–120. https://doi.org/10.1146/annurev-genet-111212-133526

Wang, Y., Tan, L., Fu, Y., Zhu, Z., Liu, F., Sun, C., Cai, H., 2015. Molecular Evolution of the Sorghum Maturity Gene Ma3. PLoS One 10. https://doi.org/10.1371/journal.pone.0124435

Wang, Y.-H., Upadhyaya, H.D., Burrell, A.M., Sahraeian, S.M.E., Klein, R.R., Klein, P.E., 2013. Genetic Structure and Linkage Disequilibrium in a Diverse, Representative Collection of the C4 Model Plant, Sorghum bicolor. G3 (Bethesda) 3, 783–793. https://doi.org/10.1534/g3.112.004861

Westengen, O.T., Okongo, M.A., Onek, L., Berg, T., Upadhyaya, H., Birkeland, S., Khalsa, S.D.K., Ring, K.H., Stenseth, N.C., Brysting, A.K., 2014. Ethnolinguistic structuring of sorghum genetic diversity in Africa and the role of local seed systems. PNAS 111, 14100–14105. https://doi.org/10.1073/pnas.1401646111

Xu, K., Xu, X., Fukao, T., Canlas, P., Maghirang-Rodriguez, R., Heuer, S., Ismail, A.M., Bailey-Serres, J., Ronald, P.C., Mackill, D.J., 2006. *Sub1A* is an ethylene-response-factor-like gene that confers submergence tolerance to rice. Nature 442, 705–708. https://doi.org/10.1038/nature04920

Yoder, J.B., Stanton-Geddes, J., Zhou, P., Briskine, R., Young, N.D., Tiffin, P., 2014. Genomic Signature of Adaptation to Climate in Medicago truncatula. Genetics 196, 1263–1275. https://doi.org/10.1534/genetics.113.159319

Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C., Weir, B.S., 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28, 3326–3328. https://doi.org/10.1093/bioinformatics/bts606

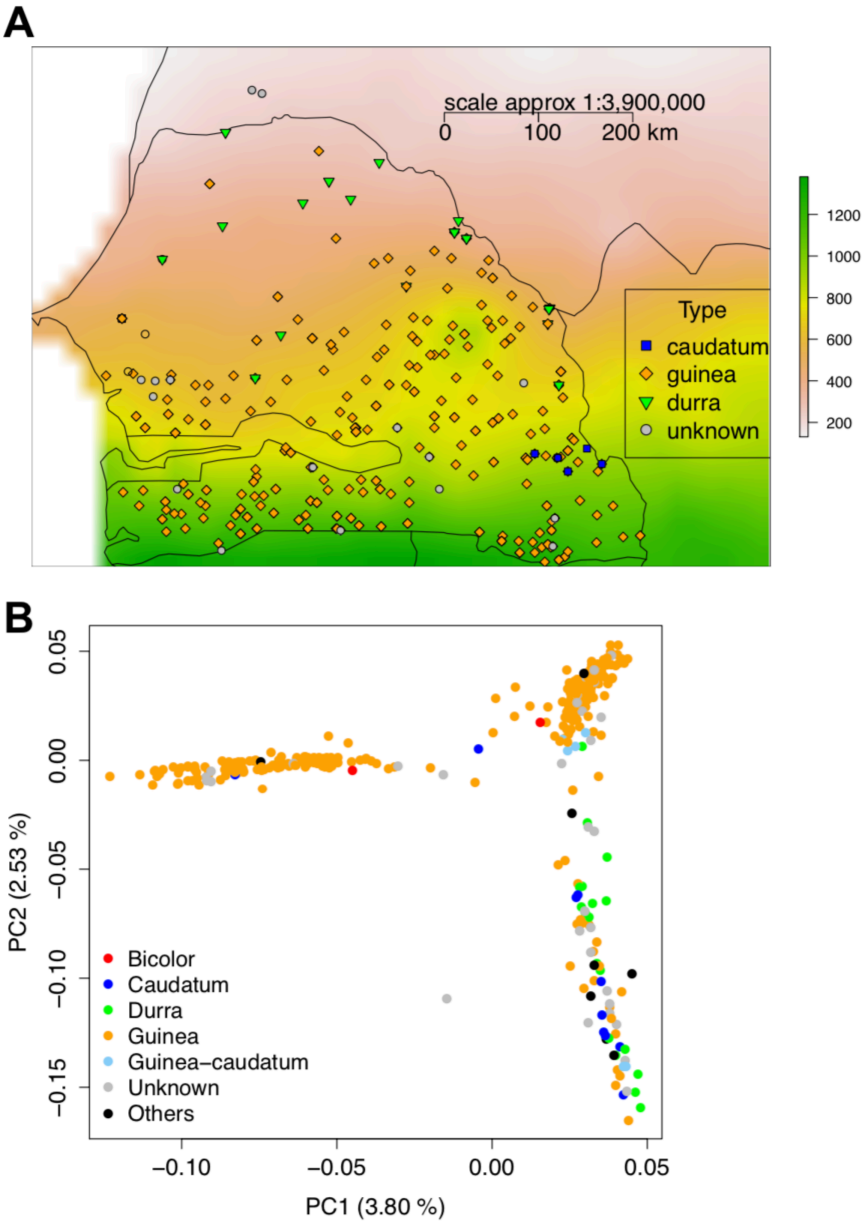



**Figure 2-1. SNP variation in the Senegalese sorghums accessions.** (A) Geographic distribution of the Senegalese sorghums accessions along precipitation gradient. The accessions are colored coded with respect to botanical race. The color background scale indicates the annual precipitation in millimeters with green color representing the highest precipitation of the Soudanian zone; pink representing lowest precipitation of the Sahelian zone, and yellow representing the zone of transition between Sahelian and Soudanian zones. (B) Scatterplot of the two first principal components explaining the genomic variation within the SSG collection.

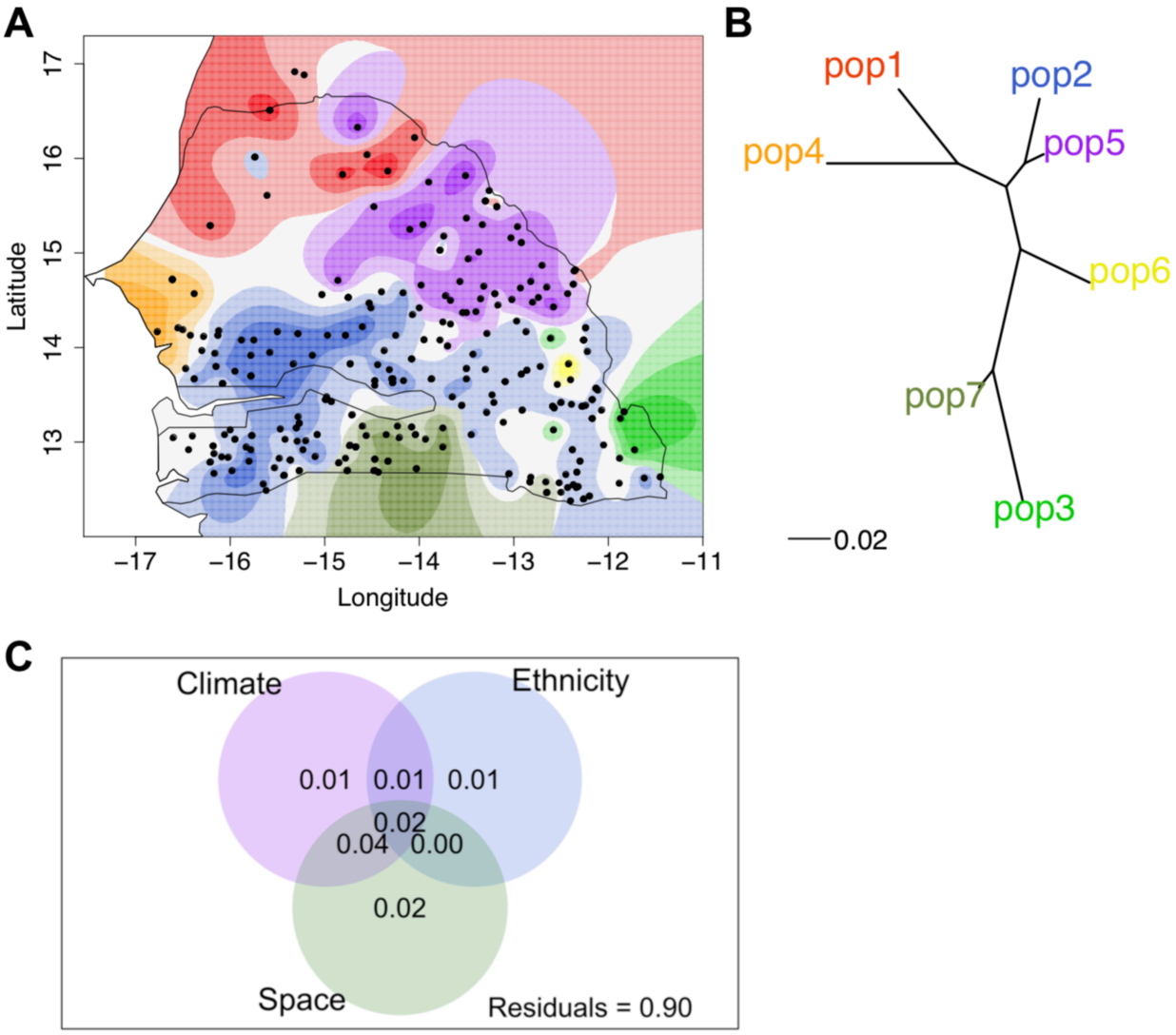



**Figure 2-2. Spatial population structure and SNP variance partitioning in the Senegalese sorghum.** (A) Spatial genetic co-ancestry structure of the accessions at K = 7. Each accession is represented by dot on the map and each color represents a genetic co-ancestry matrix. (B) The $F_{ST}$ genetic differentiation among subpopulations at K = 7 ancestral groups from B; the color-coding matches that in A. (C) Among-population genetic variance at 1000 randomly selected SNPs with MAF > 0.05 explained independently by climatic, space, and ethnicity variables.

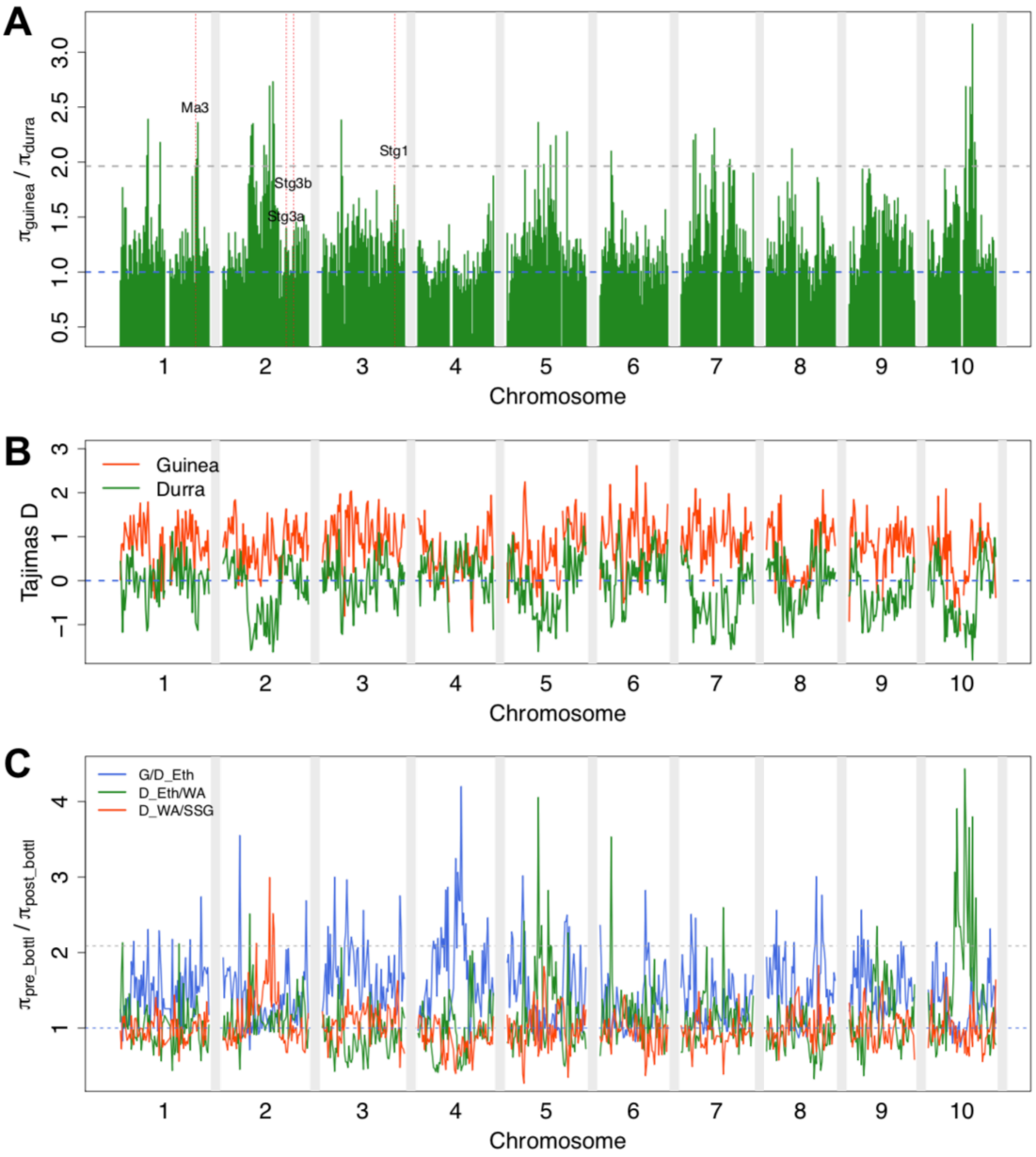



**Figure 2-3. Genome-wide pattern of nucleotide diversity in durra accessions.** Decrease in pairwise nucleotide diversity and Tajima's *D* test for non-overlapping sliding windows of 1 Mbp across the genome. (A) Decreased pairwise nucleotide diversity in durra relative to guinea in the Senegalese sorghum. The horizontal dashed lines indicate the mean value (blue) and the top 5% (gray) of decreased nucleotide diversity. (B) Tajima's *D* test between durra (green) and guinea (red) accessions in Senegalese sorghum. (C) Positive selections between durra from Ethiopia and all guineas in the global diversity panel (blue), between Ethiopian durra and West African durra (green), and between West African durra and Senegalese durra (red).

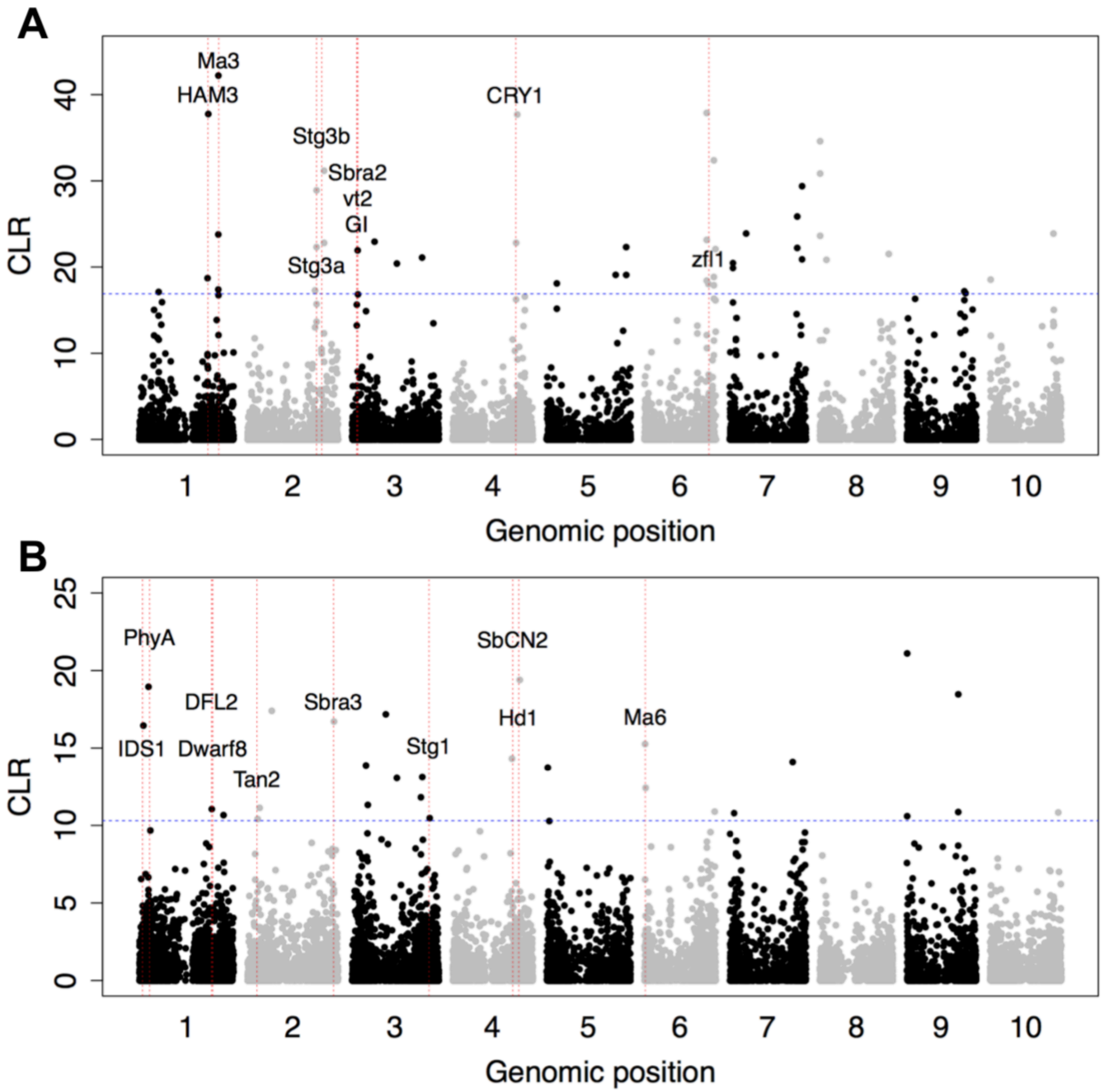



**Figure 2-4. Genome-wide scan for selective sweeps in the Senegalese sorghum.** Selective sweeps in the durra (A) and guinea (B) genomes. Each chromosome was divided into 5,000 grid points each corresponding to one dot. The y-axis represents the composite likelihood ratio (CLR) of each grid point. The vertical dashed lines indicate the co-localized candidate genes with genomic signatures. The horizontal dashed blue line represents the 95$^{th}$ percentile cutoff obtained from 1000 simulations.

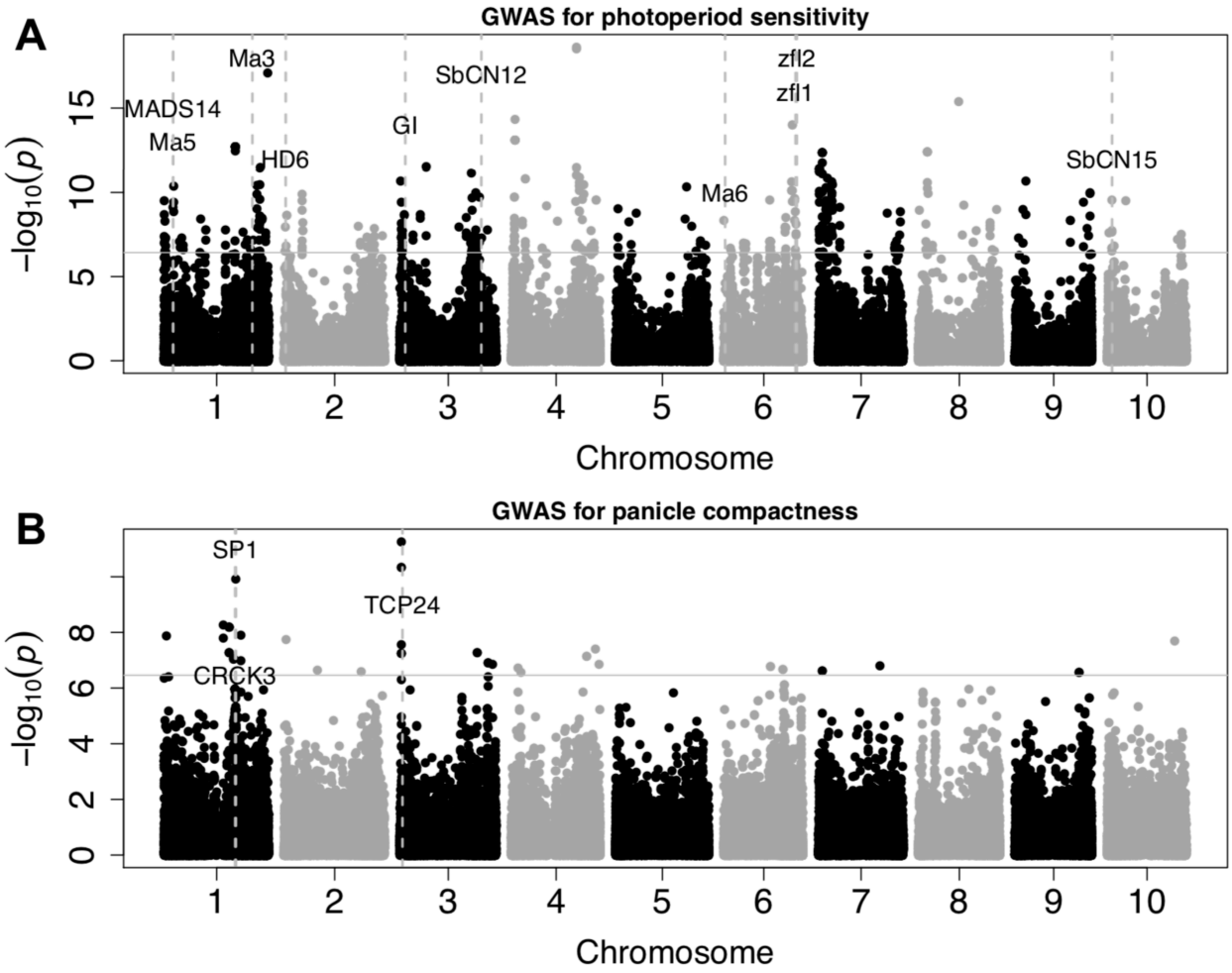



**Figure 2-5. GWAS of photoperiod sensitivity and panicle compactness.** Manhattan plots of association tests using the Mixed-linear model for photoperiod sensitivity (A) and panicle compactness (B) for the whole Senegalese collection. The negative base 10 logarithm of the significance *p*-value (y-axis) of the SNP-phenotype association is plotted against the genomic position of each SNP on the chromosomes represented on the x-axis. The gray horizontal line indicates the significance threshold for the Bonferroni corrected *p*-value > 0.05. Candidate genes co-localizing with significantly associated SNPs are indicated.

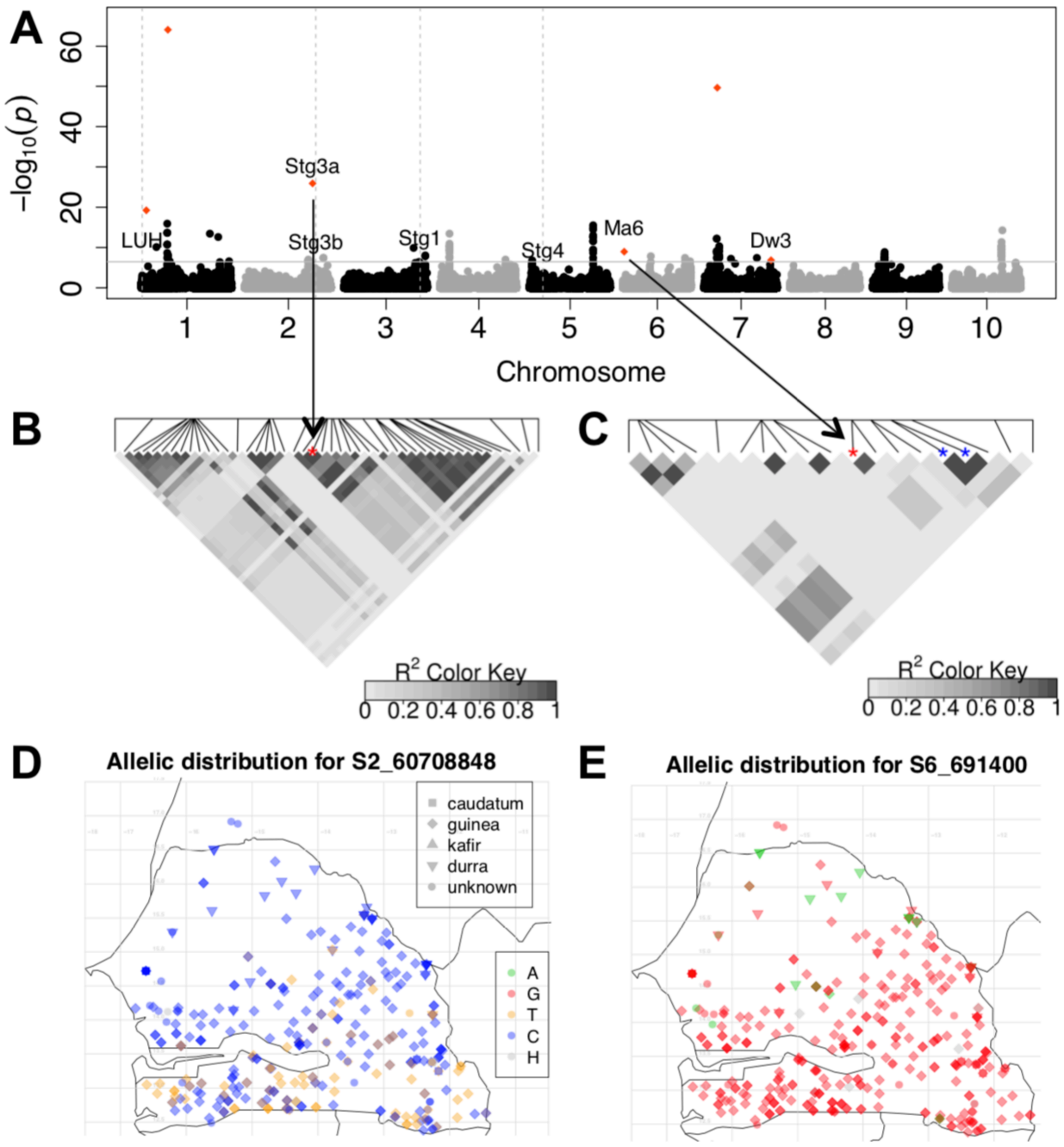

A
−log10(p)
0
20
40
60
LUH
Stg3a
Stg3b
Stg1
Stg4
Ma6
Dw3
1
2
3
4
5
6
7
8
9
10
Chromosome
B
R2 Color Key
0 0.2 0.4 0.6 0.8 1
C
R2 Color Key
0 0.2 0.4 0.6 0.8 1
D
Allelic distribution for S2_60708848
caudatum
guinea
kafir
durra
unknown
A
G
T
C
H
E
Allelic distribution for S6_691400

**Figure 2-6. Genome-environment associations for precipitation.** (A) SNP associations for "precipitation of the driest quarter" using the generalized-linear model (GLM). The red dots represent SNPs identified from the multi-locus mixed-model (MLMM). Linkage disequilibrium displayed as heat map of coefficient of correlation $r^2$ in a 50 kb region around SNPs S2_60708848 (B) and S6_691400 (C) that co-localize with *Stg3a* and *Ma6* loci in (A), respectively. Red asterisks on each heat map represent these SNPs and blue asterisks indicate the SNPs within *Ma6*. The color scale indicates the significance of $r^2$ values with black color indicating high $r^2$ values. Allelic map distribution at SNPs S2_60708848 (D) and S6_691400 (E) associated with precipitation of the driest quarter. The shape of the points indicates the botanical race of the accession and the color indicates the allele at the SNP with H being the heterozygous alleles.

# Supplemental Materials Chapter 2

This section includes the supplemental figures and tables for the chapter 2
Genomic Signatures of Adaptation to Sahelian and Soudanian Climates in Sorghum Landraces of Senegal

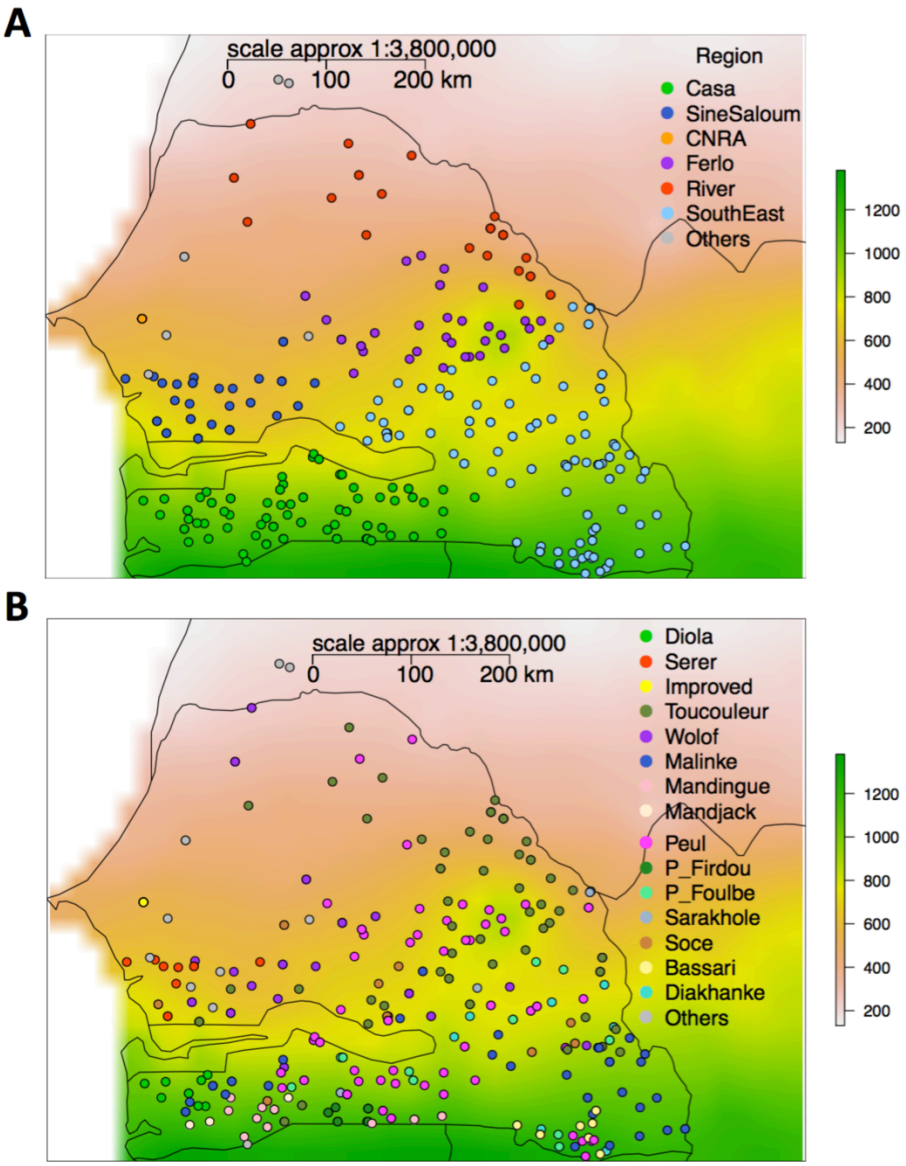



**Figure A- 1. Map of Senegalese accessions distribution colored-coded with respect to geographic region of origin (A) and ethno-linguistic groups (B).** The color background scale indicates the annual precipitation in millimeters with green color representing the highest precipitation of the Soudanian zone, pink representing lowest precipitation of the Sahelian zone, and yellow representing the zone of transition between Sahelian and Soudanian zones. The improved varieties (yellow) were assigned to the coordinates of the Centre National de Recherche Agronomic (CNRA) of Bambey, where they were developed. Casa corresponds to the region of Casamance, P_Firdou and P_Foulbe correspond to Peul Firdou and Peul Foulbe, respectively.

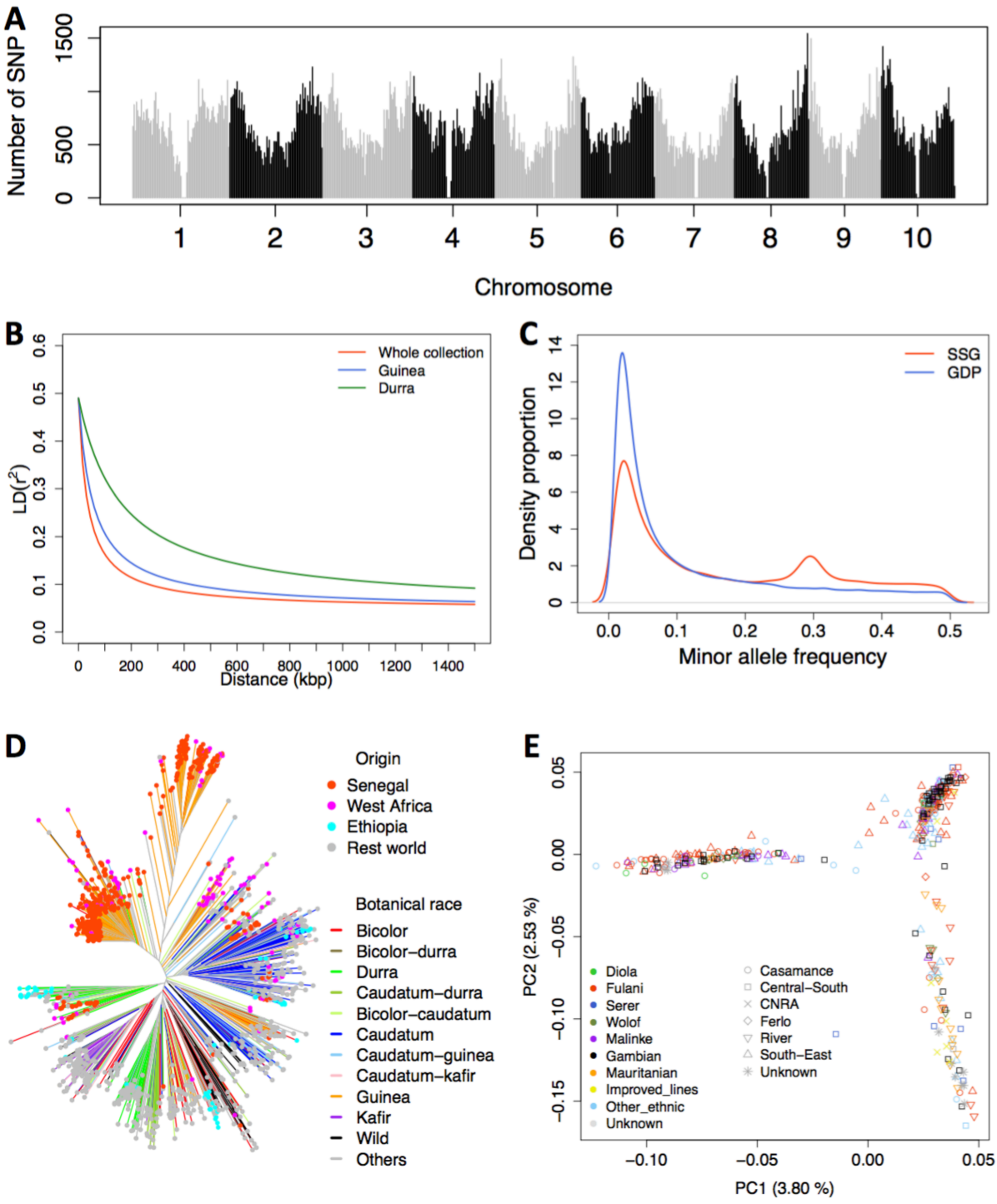



**Figure A- 2. Nucleotide polymorphisms variation and relationship between the Senegalese sorghum landraces and worldwide sorghums.** (A) SNP markers (213,916 SNPs, MAF > 1%) distribution across the 10 sorghum chromosomes of the Senegalese sorghum landraces in the USDA Germplasm Resources Information Network genebank (SSG). SNP marker density was

determined based on non-overlapping window size of 1 Mb. (B) Linkage disequilibrium decay along the genome in the whole SSG, in guinea accessions, and in durra accessions within the SSG. (C) Distribution of minor allele frequencies for the SNP data sets across the 421 accessions in the SSG (red) and 580 accessions in the global sorghum diversity panel (GDP) (blue). (D) Neighbor-joining tree based on genetic similarities between accessions of the SSG (red dots), other West African accessions (magenta dots), Ethiopian accessions (cyan dots), and the GDP (gray dots). The botanical types are represented by edges of the tree.

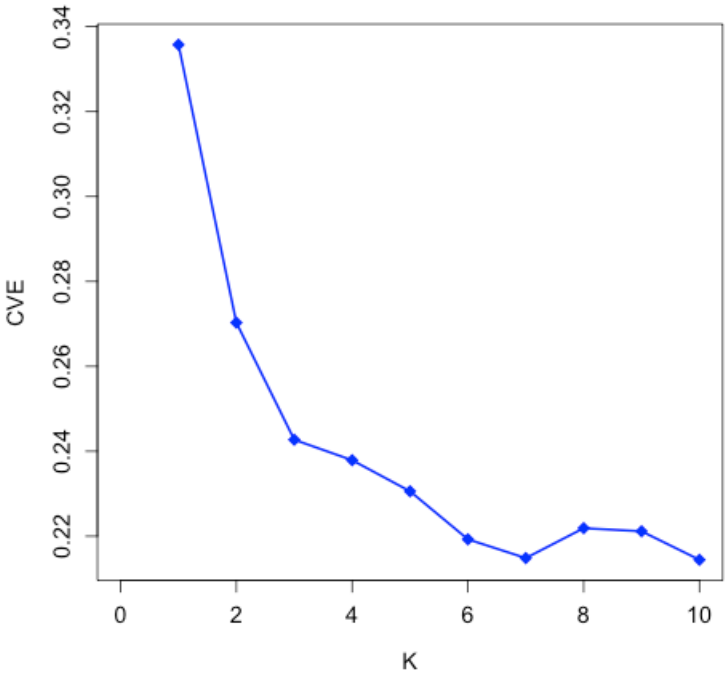



**Figure A- 3. Cross validation error of the model-based clustering of ADMIXTURE program.** The optimum number of subpopulations corresponded to K = 7.

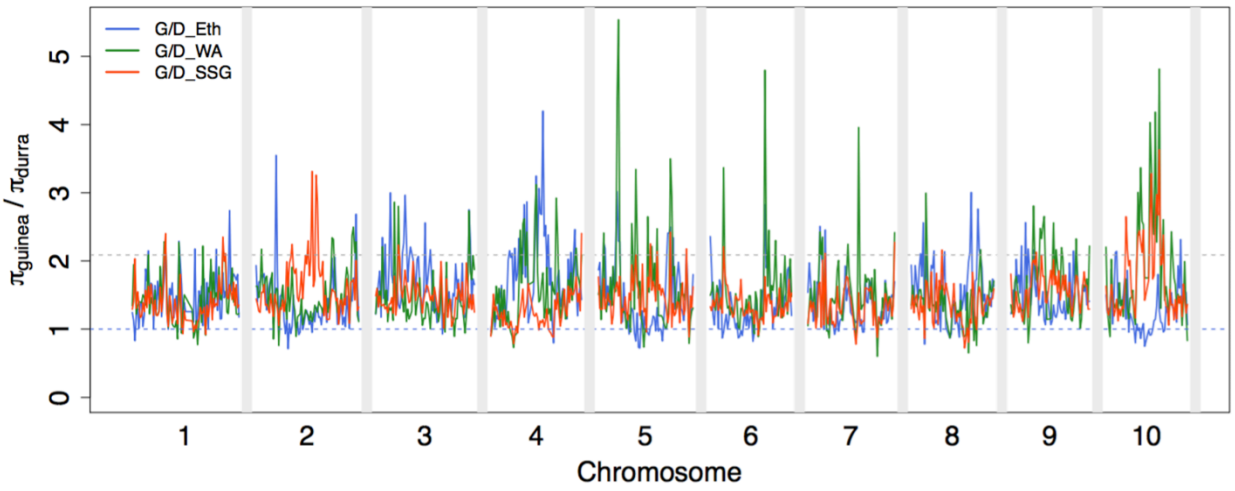



**Figure A- 4. Decreased pairwise nucleotide diversity in durra sorghums along the Sahel.** Signatures of positive selection between all guineas in the GDP and Ethiopian durra (blue), West African durra–Niger and Mali (Green), and Senegalese durra in the SSG (red). The horizontal dashed lines indicate the mean value (blue) and the top 5% (gray) of the decreased nucleotide diversity based on non-overlapping sliding windows of 1Mbp.

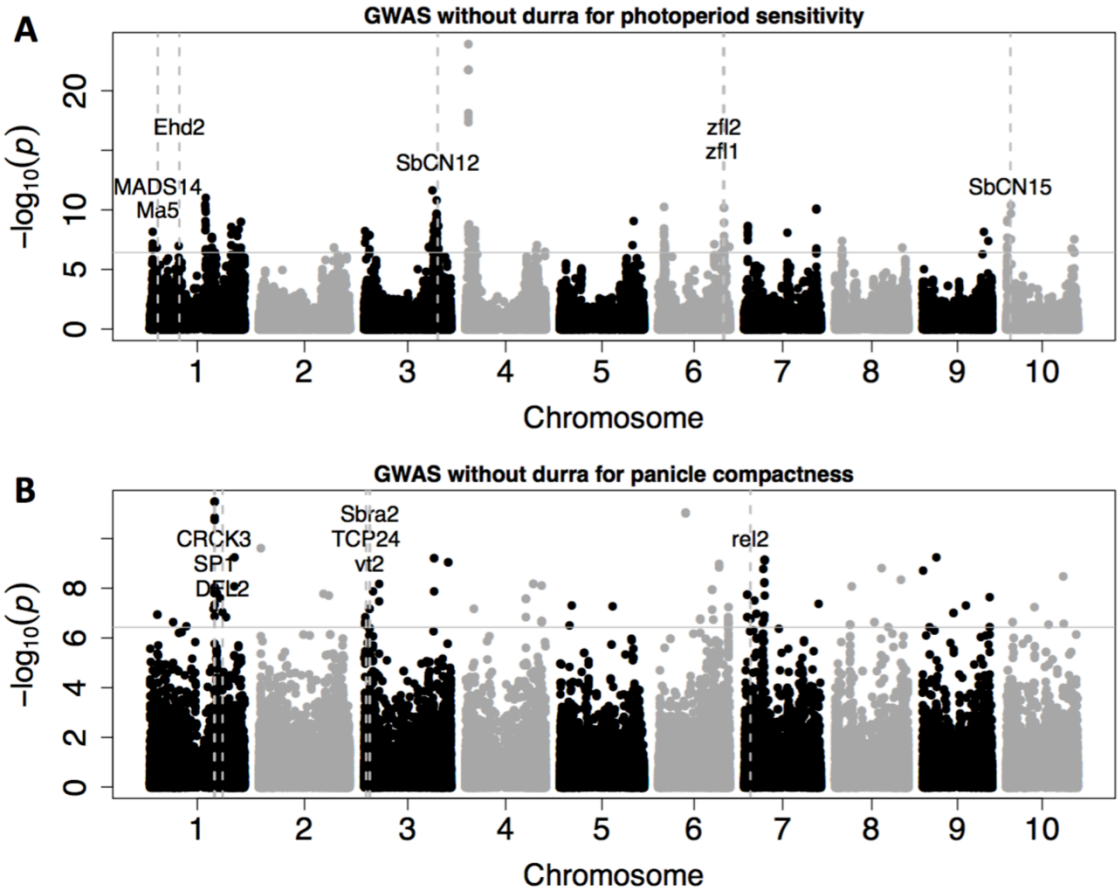



**Figure A- 5. Manhattan plots for the regional mapping GWAS, where durra sorghums accessions were excluded, using the Mixed-linear model for photoperiod sensitivity (A) and panicle compactness (B) for the SSG sorghum landraces.** The negative base 10 logarithm of the significance *p*-value (y-axis) of the SNP-phenotype association is plotted against the genomic position of each SNP on the ten chromosomes (x-axis). The gray horizontal line indicates the significance threshold for the Bonferroni corrected *p*-value > 0.05. Candidate genes co-localizing with significantly associated SNPs are indicated.

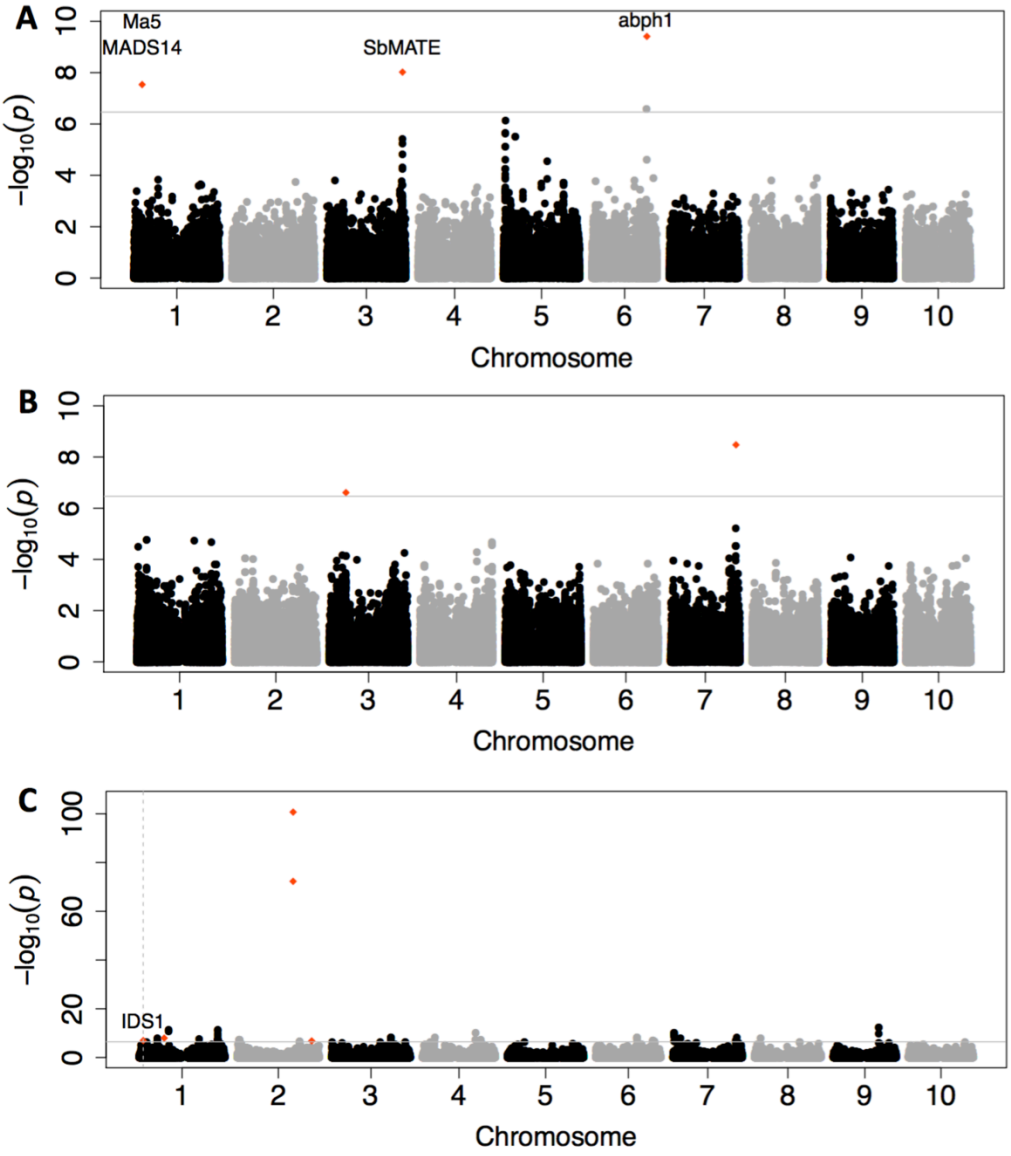



**Figure A- 6. Genotype-environment associations for adaptation in the Senegalese sorghum landraces.** (A) Manhattan plot showing SNPs associated with mean temperature of the warmest quarter using the generalized-linear model (GLM). (B) SNPs associated with precipitation of the wettest quarter using GLM. (C) Manhattan plot showing SNPs associated with longitude using the mixed-linear model (MLM). The GLM identified many associations for longitude variable, so the MLM results are showed here. The red dots on each plot represent significantly associated SNPs identified from the multi-locus mixed-model (MLMM). The x-axis represents the SNP position on the ten chromosomes of sorghum. The y-axis indicates the significance of the associations.

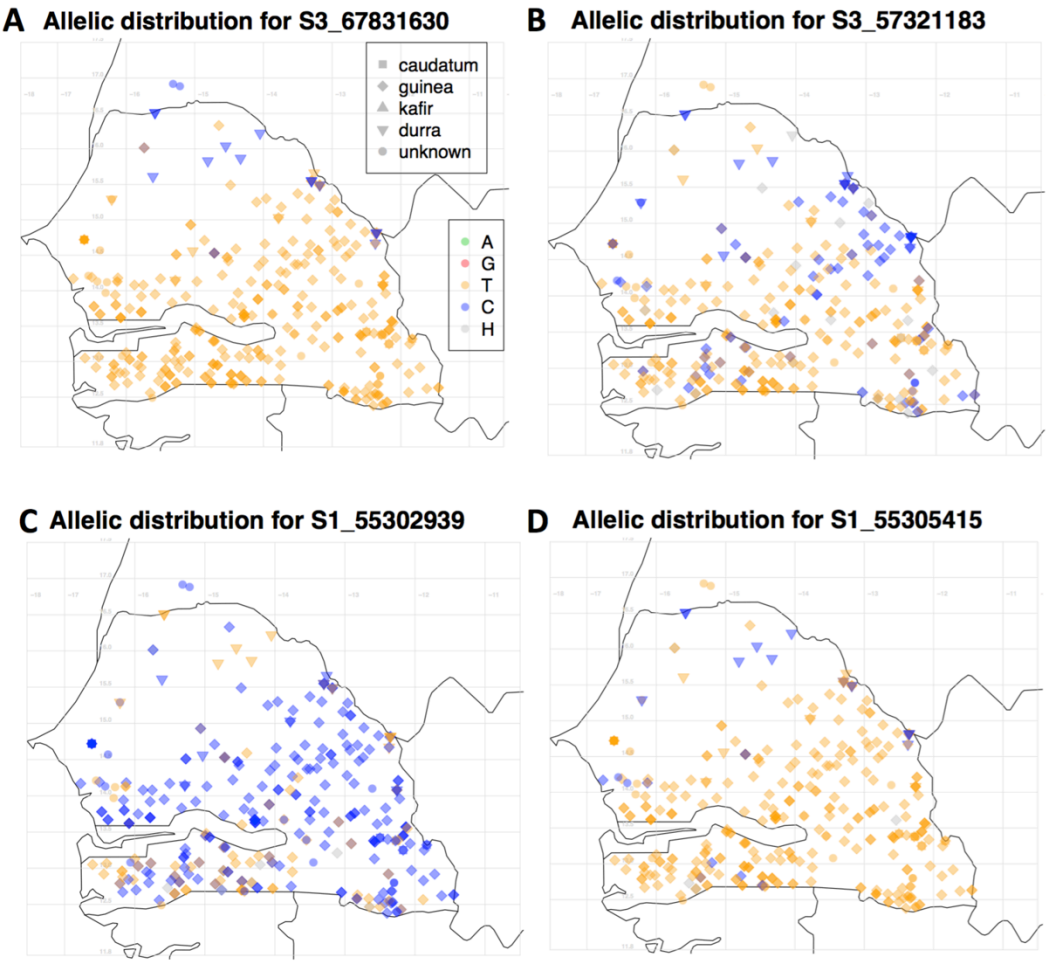



**Figure A- 7. Allele distribution along the map of Senegal.** (A) Allelic map distribution at SNP S3_67831630 significantly associated with precipitation of the driest quarter and co-localizing with *Stg1/SbPIN4* locus. The minor allele at this SNP is found in durra landraces distributed in the dry areas in the Sahelian zone of Senegal. (B) Allelic map distribution at SNP S3_57321183 associated with precipitation of the wettest quarter and co-localizing with *Stg2/SbPIN2* locus. The minor allele at this SNP is mostly found in durra and a few guinea accessions of in the drier areas of Senegal. Allelic map distribution at SNPs S1_55302939 (C) and S1_55305415 (D) associated with panicle compactness and co-localizing with the *SP1* candidate gene.

**Table A- 1. SNP associated with environment variables using the multi-locus mixed-linear model (MLMM).** In the SNP column, the digit after "S" indicates the chromosome number and the other digits after the underscore indicate the SNP position.

| SNP | P-value | MAF | Closest gene/locus | Position to gene (kb) |
|---|---|---|---|---|
| *Precipitation of the driest quarter* | | | | |
| S1_23271724 | 7.87E-65 | 0.287323944 | | |
| S7_11301809 | 2.23E-50 | 0.270422535 | | |
| S2_60708848 | 1.25E-26 | 0.281690141 | *Stg3a* | within |
| S1_3836265 | 5.85E-20 | 0.285915493 | | |
| S6_691400 | 1.08E-09 | 0.073239437 | *Ma6* | 6 |
| S7_59683060 | 1.40E-07 | 0.050704225 | *Dw3* | 138 |
| *Mean temperature of the warmest quarter* | | | | |
| S1_7584419 | 2.92E-08 | 0.090140845 | MADS14; *Ma5* | 884; 830 |
| S6_51709806 | 3.89E-10 | 0.243661972 | *abph1* | 454 |
| S3_71370900 | 9.54E-09 | 0.173239437 | *SbMATE* | 262 |
| *Precipitation of the wettest quarter* | | | | |
| S7_61856992 | 3.34E-09 | 0.091549296 | | |
| S3_15688290 | 2.46E-07 | 0.333802817 | | |
| *Longitude* | | | | |
| S2_52993994 | 2.05E-101 | 0.057746479 | | |
| S2_52937592 | 5.40E-73 | 0.057746479 | | |
| S1_23473920 | 1.07E-08 | 0.290140845 | | |
| S1_3547581 | 1.11E-07 | 0.057746479 | *IDS1* | 806 |
| S2_70618491 | 1.56E-07 | 0.285915493 | | |

# Chapter 3 - A Genomic Resource for Genetics, Physiology, and Breeding of West African Sorghum

A chapter to be submitted to The Plant Genome as:
Jacques M. Faye, Fanna Maina, Eyanawa A. Akata, Bassirou Sine, Niaba Teme, Cyril Diatta, Aissata Mamadou, Sandeep Marla, Sophie Bouchet, Jean-Francois Rami, Daniel Fonceka, Ndiaga Cisse, Geoffrey P. Morris. A Genomic Resource for Genetics, Physiology, and Breeding of West African Sorghum.

**Abbreviations**

BLUP, best linear unbiased prediction; CVE, cross-validation error; DFLo, days to flowering; GBS, genotyping-by-sequencing; GDP, global sorghum diversity panel; GLM, general linear model; GWAS, genome-wide association studies; LD, linkage disequilibrium; MLMM, multi-locus mixed linear model; PH, plant height; PVE, proportion of phenotypic variance explained; PW, panicle weight; QTL, quantitative trait loci; SNP, single nucleotide polymorphism; USDA-GRIN, United States Department of Agriculture, Germplasm Resources Information Network; WA, West Africa; WASAP, West African sorghum association panel.

## Abstract

Local sorghum (*Sorghum bicolor*) germplasm lines are useful sources of genetic diversity and adaptive traits for efficient crop improvement. Genetic characterization of a new collection of West African sorghums would facilitate their use in genome-wide studies and breeding. In this study, a West African sorghum association panel (WASAP) of 756 accessions was assembled from breeding programs of Senegal, Mali, Togo, and Niger. Genotyping-by sequencing was used to generate 159,101 high-quality biallelic SNPs with < 20% missing data and minor allele frequency > 0.01 across accessions. About 43% of SNPs were mapped in intergenic regions and 13% in genic regions, including missense (72%), nonsense (4%), and silent (24%) point mutations. High genetic diversity was observed within the WASAP ($\pi = 0.00045$), only slightly less than in a global diversity panel ($\pi = 0.00055$). Diversity structured by botanical type and subpopulation within botanical type across countries. Linkage disequilibrium decayed to background level ($r^2 < 0.1$) by ~50 kb in the WASAP. Eight ancestral populations were identified, and clustered along with WA accessions in USDA-GRIN and the global diversity

panel. GWAS revealed eight and three significantly associated quantitative trait loci (QTLs) with days to flowering and plant height, respectively. Several large effect QTLs for flowering time were colocalized with known genes *Ma6* and *SbCN8*, and several novel loci, indicating oligogenic architecture for flowering time with *Ma6* being the potential flowering time gene in the WA sorghum. Altogether, this study provides genetic and genomic resources for efficient utilization of the WASAP for developing climate resilient varieties.

## Introduction

Crop production in many developing countries is limited by biotic and abiotic factors that reduce food supplies to smallholder farmers in semi-arid areas. With an increasing worldwide underfed population along with environmental changes, there is a need in more rapidly developing locally adapted varieties to increase crop productivity (Foley *et al.*, 2011; Tilman *et al.*, 2011; Mundia *et al.*, 2019). Genetic studies contribute to the development of adapted varieties to meet global food security and help provide enough genetic diversity suitable for efficient crop breeding (Jordan *et al.*, 2011). Diverse landrace germplasm harbors useful alleles for gene discovery and breeding, given their long history of adaptation to diverse environments (Meyer & Purugganan, 2013). However, African crop genetic diversity, particularly in West Africa (WA), are poorly characterized mainly due to lack of genomic resources and limited sampling of genetic resources available to the global scientific community.

Understanding genomic variation of local germplasm at a regional scale can help guide breeding. The availability of high-density markers evenly distributed throughout the genome is a prerequisite for understanding genetic diversity and genetic basis of adaptive traits. Recent advances in next-generation sequencing technologies and GBS have rendered possible the generation of high-density markers with affordable low cost (Elshire *et al.*, 2011; Poland *et al.*, 2012). These tools facilitate characterization of the genetic structure of local germplasm relative to global diversity. Historical recombination along with short to moderate linkage disequilibrium (LD) existing within a diversity panel greatly improve the mapping resolution to identify novel genes and novel natural variants at known genes in major crops (Huang *et al.*, 2010; Yano *et al.*, 2016; Cao *et al.*, 2016; Gapare *et al.*, 2017; Zhao *et al.*, 2019).

In sorghum, global reference diversity panels have been assembled (Grenier *et al.*, 2001; Deu *et al.*, 2006; Casa *et al.*, 2008; Upadhyaya *et al.*, 2009; Billot *et al.*, 2013; Brenton *et al.*, 2016) and used in genetic studies and breeding. However, an underrepresentation of accessions

at regional scale arises in these reference panels due to limited population sampling, thus limiting their potential use in regional association mapping studies. Regional diversity panels are useful genetic resources for capturing natural allelic variation existing in locally-adapted varieties (Leiser *et al.*, 2014; Sattler *et al.*, 2018). Favorable alleles for adaptation to various regional environmental conditions have been selected for over thousands of years; however, they might be lost over time at different levels–farmers practices, modern varieties breeding and diffusion, and *ex situ* collections (Hammer & Teklu, 2008; Fu, 2017). The major West and Central African cereal crops such as sorghum and pearl millet are poorly assembled into regional reference diversity panels suitable for genetic studies and breeding (Sattler *et al.*, 2018).

Genetic diversity of cultivated sorghum is high in West African (Doggett, 1988; Deu *et al.*, 1994; Folkertsma *et al.*, 2005). Despite the high genetic diversity in the West African germplasm, its accessibility to the regional and global scientific community is limited, particularly germplasm that was more recently collected or developed. Five botanical types in sorghum—bicolor, durra, guinea, caudatum and kafir and ten intermediate types—have been defined based on spikelet and grain morphology (Harlan & De Wet, 1972) and are associated with climate and geographic origin (Brown *et al.*, 2011). All botanical types are represented in West African landraces except for the kafir type. The guinea type is the most common and diverse in WA, possibly due to a second center of domestication (Folkertsma *et al.*, 2005). The durra type was first domesticated in Ethiopia before diffusing to arid regions of West Africa. The caudatum and durra-caudatum intermediate types are well represented in the west-central region, and used for grain yield improvement in breeding programs throughout West Africa (ISRA, 2005). However, little is known about the population structure of the germplasm across the countries, whether the germplasm of one country is distinct from other countries. Since country boundaries in West Africa cut across agro-ecological zones, some landrace germplasm may be more similar across countries than within a country.

Here, we report the assembly of the WASAP from the four West African countries (Mali, Niger, Senegal and Togo) and development of genome-wide SNP markers as genomic resources for genetics, physiology, and breeding. We determined the genome-wide SNP variation and population structure of the germplasm in relationship with previously genotyped West African *ex situ* collections and global sorghum diversity panel (GDP). We also identified known maturity and plant height loci using GWAS on phenotype data collected under rainfed conditions. The

study provides genomic resources and a better understanding of the population structure of the WA germplasm useful for genomics-assisted breeding.

## Materials and Methods

### Plant materials

The WASAP was composed of 756 accessions assembled by breeders, physiologists, and geneticists in national agricultural research organizations (NARO) from four West African countries (Senegal, Mali, Togo, Niger) (Table 3-1; Supplemental Data S1). The panel includes working collections of the NARO sorghum improvement programs, predominantly landraces, but also locally-improved varieties and breeding lines, and well represents all four basic botanical types (bicolor, caudatum, durra, and guinea), except the kafir type (n = 1), based on *a priori* classification. Many accessions were not classified morphologically into botanical classes (n = 230). Genetic diversity and structure of the WASAP were compared with the global diversity panel (GDP) that consists of 692 worldwide sorghum accessions (excluding accessions from Americas) with available sequencing data, including West African accessions in the GDP (hereafter named WAS-GDP) (Morris *et al.*, 2013; Lasky *et al.*, 2015). WASAP accessions were also compared to previously genotyped West African accessions from USDA-GRIN (hereafter named WAS-GRIN), originating from Niger, Senegal (including a few accessions from neighboring Gambia and Mauritania), and Nigeria (Maina *et al.*, 2018; Olatoye *et al.*, 2018; Faye *et al.*, 2019).

### Genotyping-by-sequencing and SNP discovery

The WASAP was grown in the field at the Bambey research station in Senegal and leave tissue from five seedlings of each accession were pooled to extract genomic DNA using the MATAB (Mixed Alkyl trimethylammonium bromide) protocol. Genotyping-by-sequencing (GBS) was conducted following the method previously described (Elshire *et al.*, 2011). Briefly, GBS libraries were constructed in 96-plex and the restriction enzyme, *Ape*KI (New England Biolabs), was used to digest genomic DNA for complexity reduction. For quality control, a random well was left blank in each 96-well plate. Restriction cutting sites were ligated using barcoded adapters and ligated products were pooled together for sequencing. Single-end sequencing was performed on Illumina HiSeq2500 at the University of Kansas Medical Center (Kansas, USA).

Illumina single-end sequence reads of the WASAP and raw sequence data of the GDP were processed together using the TASSEL GBS v2 pipeline (Glaubitz *et al.*, 2014). Sequence reads were trimmed to 64-bp and identical reads collapsed into tags. Tags were aligned to the sorghum reference genome v3.1 (Paterson *et al.*, 2009; McCormick *et al.*, 2018) using the Burrows-Wheeler Alignment (BWA) program (Li & Durbin, 2009). Single nucleotide polymorphism (SNP) markers were discovered using the *DiscoverySNPCallerPluginV2* of the TASSEL GBS v2 pipeline with *minimum locus coverage* (*mnLov*) of 0.1 and other parameters kept to default settings. A total of 546,133 SNPs were discovered. SNPs with more than 20% missing data were excluded (n = 393,396 remaining SNPs). SNPs with minor allele frequency (MAF) < 0.01 were excluded (n = 201,193 remaining SNPs). Monomorphic sites were excluded and biallelic SNPs only were maintained, resulting in a total set of 198,402 SNPs. This dataset was imputed using Beagle v4.1 (Browning & Browning, 2016) and filtered out again data with MAF < 0.01 to yield a final data set of 159,101 SNPs that was used for downstream analysis.

## Linkage disequilibrium analysis

Linkage disequilibrium (LD) was determined for the entire WASAP and for each country's germplasm separately. LD was estimated based on the pairwise correlation coefficient ($r^2$) among SNPs with MAF > 0.05 (to reduce computational burden) in a window of 500 kb using the PopLDdecay package (Zhang *et al.*, 2019). The *smooth.spline* function in the R program (R Core Team, 2016) was used to fit LD decay measured as the distance by which $r^2$ decreased to half from its original value.

## Genome-wide SNP variation and genetic structure analysis

The imputed dataset was functionally annotated to determine SNP effects on protein-coding genes using the snpEff program (Cingolani *et al.*, 2012). The structural location and functional class (synonymous, missense, or nonsense) of each SNP were determined based on the sorghum reference sequence. Genome-wide SNP distribution along chromosomes, minor allele frequencies, and pairwise nucleotide diversity were estimated using the VCFtools program (Danecek *et al.*, 2011). $F_{ST}$ genetic differentiation in the WASAP was determined according to botanical type and country of origin using the pairwise $F_{ST}$ method based on Weir and Cockerham weighted $F_{ST}$ estimate in the VCFtools.

The genome-wide SNP variation of the collection was assessed based on the principal component analysis (PCA) using the *snpgdsPCA* function of the R package SNPRelate (Zheng *et*

*al.*, 2012). The accessions of WAS-GRIN and GDP were included in the analysis to robustly determine principal axes of genome-wide SNP variation and clustering of the WASAP along other germplasms. The combined dataset consisted of a subset of 103,871 (100K) high-quality SNP markers with a high level of polymorphism (MAF > 0.1). The GDP accessions were used as a training set to calculate principal components and predict the genetic structure among the WASAP and WAS-GRIN accessions.

The number of ancestral populations and ancestry fractions in the WASAP were determined using ADMIXTURE (Alexander *et al.*, 2009). The original SNP dataset was LD-pruned to generate 60,749 SNPs using the PLINK 1.9 (Purcell *et al.*, 2007) to reduce redundancy of SNPs that are in high LD because such SNPs provide the same genetic information. Ancestral populations and ancestry fractions were determined using the pruned SNP dataset (60,749 independent SNPs) for K = 2–20 populations using default settings of ADMIXTURE. A five-fold cross-validation and 2,000 iterations were performed to identify the optimal K. We defined the optimal K as the minimum K where cross-validation error no longer decreased substantially. Accessions were assigned to subpopulations based on 0.7 membership threshold. The neighbor-joining distance-based clustering method was used to assess the genetic relationship among WASAP ancestral populations, WAS-GRIN, and GDP. The genetic distance matrix was generated based on the 100K SNPs using TASSEL 5 (Bradbury *et al.*, 2007) and the results were plotted using *ape* R package (Paradis *et al.*, 2004).

## Field phenotyping

Field phenotyping of a subset of 572 WASAP accessions (based on seed availability) was conducted at the Bambey Research Station, Centre National de Recherches Agronomiques (14.42˚N, 16.28˚W) in Senegal during the growing season of 2014. Two experiments, early-sown date at the beginning of the rainy season–normal planting date (Hiv1) and late-sown post-flowering drought-stressed–planted 30 days later (Hiv2), were performed. Each experiment had one replication. A randomized incomplete block design (augmented block design) was performed with 30 blocks of 24 entries each, following a column-row field layout for spatial data analysis. Each block contained 19 genotypes and 5 check varieties. The five checks were randomly assigned into each block. Each genotype was assigned only once in each experiment. Each genotype was planted in one row of 3 m surrounded by one row of fill material on both sides. There was 60 cm space between rows and 20 cm space between plants or hills within a

row. Ten days after planting, genotypes were pruned to maintain only one plant per hill. Days to flowering (DFLo) was measured as the day when 50% of plants within a plot flowered. Plant height (PH) was measured as the average distance from the soil to the tip of the panicle of three plants per plot.

## Statistical analysis of phenotypic data

Phenotypic variation was analyzed using the R program (R Core Team, 2016). The genotype adjusted mean value was determined after correcting for significant spatial variation effect using the check varieties. The variance components were estimated by fitting the mixed linear model with random effects for all genotypes (G), environment (E), and G x E interaction effects using the *lme4* package (Bates et 2010). Broad sense heritability ($H^2$) was estimated for each trait from the estimated genetic and residual variances derived from the mixed effect model, as follows:

$$H^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_{GxE}^2/r + \sigma_e^2/r}$$

where $\sigma_G^2$ is the genotypic variance, $\sigma_{GxE}^2$ is the genotype by environment interaction variance, $\sigma_e^2$ is the residual error variance, and *r* the number of environments. The early and late planting date experiments were considered as two environments. Phenotypic correlations among traits were calculated using the Pearson correlation of the *PerformanceAnalytics* package (Peterson *et al.*, 2014). BLUP values were calculated by combining data from environments. Tukey's Honestly Significant Difference (TukeyHSD) in the Agricolae package (Mendiburu, 2009) was used to test the difference of genotype performance between environments.

## Genome-wide association studies

Genome-wide association studies (GWAS) were performed using the general-linear model (GLM) in the GAPIT R package (Lipka *et al.*, 2012) and multi-locus mixed-model (MLMM) in the *mlmm* package (Segura *et al.*, 2012). The MLMM stepwise regression model was used to account for background effects. The GLM does not account for background effects but shows most significant associations. In contrast, the mixed-model method implemented in MLMM accounts for background effects but increases false negative associations. These different methods were used as complementary. We did not consider looking at nominal p-value as we are aware that the GLM results are generally inflated. Our goal was to test the hypothesis that some of the top associations would colocalize with known candidate genes.

A total of 130,709 SNPs with MAF > 0.02 were used for the GWAS analysis, as moderately rare variants can contribute to phenotypic variation (Peloso *et al*., 2016; Hernandez *et al*., 2019). The first three principal components generated from TASSEL 5 (Bradbury *et al.*, 2007) and kinship matrix from GAPIT were used to account for polygenic background effects in the MLMM analysis. GWAS were performed for BLUP values of DFLo and PH across two environments. LD heatmaps of genomic regions surrounding GWAS QTL were constructed using the package *LD heatmap 0.99-4* package (Shin *et al.*, 2006). All figures were produced with the R program (R Core Team, 2016). The effect size and proportion of phenotypic variance explained by associated QTLs were determined using linear regression and ANOVA. Background effect was accounted for using ADMIXTURE ancestry fractions used as fixed effect covariates. Candidate gene colocalization with QTL was carried out using an *a priori* candidate genes/loci list, including known sorghum genes and orthologs of rice and maize for adaptive traits, previously described (Faye *et al.*, 2019). Because candidate genes were *a priori* defined, we used a liberal cutoff of 500 kb to determine colocalization between association signals and candidate genes. The Sorghum QTL Atlas (Mace *et al.*, 2019) was used to compare QTLs identified in the current study to QTLs from previous studies.

## Results

### Genome-wide SNP variation of the West African sorghum association panel

The GBS library sequencing yielded a total of ~258 million single-end sequencing barcoded reads. After trimming all reads down to 64 bp, ~4.5 million unique tags were obtained. A final data set of 159,101 high-quality SNPs was maintained after removing SNPs with >20% missing data, MAF < 0.01, and keeping only biallelic SNPs. The SNPs were distributed across the genome with a higher number of SNPs in pericentromeric regions relative to centromeric regions (Fig. 3-1A). We determined if the 159,101 GBS-SNPs have potential impacts on protein-coding sequences based on the sorghum reference sequence v.3.1. About 13% of the SNPs were found in genic regions, including 7,689 missense (72%), 411 (4%) nonsense, and 2,603 (24%) silent point mutations (Fig. B-1). About 43%, 22%, and 22% variants were located in intergenic, downstream, and upstream regions of genes, respectively.

Since four of five sorghum botanical types are represented in the WASAP, we hypothesized that it captures much of the genetic diversity found in the GDP. The average pairwise nucleotide diversity ($\pi$) was less in the WASAP (0.00045) than in the GDP (0.00055).

Little variation in π was observed among the four countries of origin (Niger: 0.00046, Mali: 0.00049, Senegal: 0.00050, Togo: 0.00047). The WASAP had a higher proportion of rare alleles (e.g., 0.01–0.05) than the GDP (Fig. 3-1B). Within the WASAP, the Senegal accessions had the lowest rare allele proportion and the highest intermediate allele proportion followed by Togo and Mali accessions. Linkage disequilibrium (LD) decayed to background level ($r^2 < 0.1$) by ~50 kb in the WASAP versus ~15 kb in the GDP (Fig. 3-1C). As expected, LD decay within countries of origin in the WASAP was higher than that in the whole WASAP. LD decayed to background by ~60 kb, ~90 kb, ~90 kb, and ~160 kb in Niger, Senegal, Mali, and Togo accessions, respectively.

## Genetic differentiation by botanical types and geographic origin

Based on the hypothesis that sorghum genetic diversity is structured primarily by botanical type, we predicted high $F_{ST}$ genetic differentiation would be observed among botanical types than among countries of origin in the WASAP. High $F_{ST}$ genetic differentiation of 0.16 was observed among the six classes composing the majority of the panel, including guinea, caudatum, durra, bicolor, guinea-margaritiferum (Gm) types, and the intermediate form durra-caudatum (DC) (Supplemental Table S1). Surprisingly, high $F_{ST}$ value was observed between DC and caudatum ($F_{ST} = 0.22$) or DC and durra ($F_{ST} = 0.20$). The $F_{ST}$ value among the four countries of origin was moderate ($F_{ST} = 0.09$) (Table B-1).

We characterized the genomic variation of the WASAP along with the WAS-GRIN and GDP. The first two PCs explained a high proportion of genome-wide SNP variation (a combined 17%) and differentiated the caudatum, durra, guinea, and kafir accessions (Fig. 3-2A). We predicted that the majority of the WASAP accessions will be clustered with guinea, caudatum, and durra accessions from the same geographic origin in the GDP. The majority of the WASAP accessions overlapped with their corresponding types, guinea, durra, and caudatum clusters of GDP along the PC2. A substantial variation was explained by both PC3 vs. PC4 (7.3%) (Fig. 3-2B). The durra-caudatum intermediate types in the WASAP and WAS-GRIN clustered between durra, caudatum, and guinea clusters in the GDP.

## Ancestral fractions and population structure

To determine the ancestral populations and ancestry fractions for each accession, we used the Bayesian model-based method ADMIXTURE. Based on five-fold cross-validation error (CVE), the optimum number of ancestral populations was eight (Fig. 3-3A). The accessions classified morphologically as guinea (orange lower rug-plot) corresponded to three genetic

groups (G-II, G-IV, G-VII) (Fig. 3-3B). The accessions classified morphological as durra-caudatum intermediates (mostly from Niger; green lower rug-plot) corresponded to two genetic groups (G-V, G-VIII). Accessions classified morphologically as caudatum (blue lower rug-plot) corresponded to G-I. Using 0.7 ancestry fraction as a threshold, 71% of accessions could be assigned to a subpopulation, while 29% would be considered admixed. The greatest putative contribution to genetic admixture was from G-V (purple bars), with ancestry fraction present in all other subpopulations. $F_{ST}$ among ancestral populations averaged 0.39, with a range of 0.25–0.61(Table B-2).

Next, we considered the extent to which germplasm of each country is distinct from the germplasm of other countries (Fig. 3-3C). Each of the countries' germplasm included multiple genetic groups. Most of the ancestral populations were found in each country, except in Togo where only three genetic subpopulations (G-I, G-II, G-VII) were clearly defined. The G-IV was specific to Senegal and Mali, while G-VI was specific to Niger and Mali. G-VII and G-VIII groups were specific to Togo and Niger, respectively. Neighbor-joining (NJ) analysis recapitulated the country-level ancestry structure (see color-coded tips) (Fig. 3-4 and Fig. B-2). Genetic similarities were observed between WASAP ADMIXTURE genetic groups and other West African sorghums (WAS-GRIN and WAS-GDP) from the same geographic origin. and GDP accessions according to botanical type and geographic origin. The West African sorghums clustered with their corresponding types in the GDP. The guinea and durra-caudatum accessions of West Africa clustered generally distinctly from the majority of GDP accessions.

## Phenotypic variation in the WASAP

A total of 572 accessions of the WASAP and five check lines were evaluated for agronomic traits under rainfed conditions. We hypothesized that the phenotypic variation in the WASAP is due to by genetic effect, appropriate for dissecting agronomic traits using GWAS. Large phenotypic variation was observed for both DFLo (21%) and PH (35%). Significant genotypic (G) variation and genotype by environment (G x E) interaction effects were observed (Table 3-2). High broad-sense heritability ($H^2$) was observed across the two environments, with values ranging from 0.74 for PH to 0.88 for DFLo (Table 3-2). DFLo were significantly higher in Hiv1 (TukeyHSD $p < 0.0001$) (Fig. B-3). DFLo was significantly correlated between Hiv1 and Hiv2 ($r^2 = 0.75$, $p < 0.0001$) (Fig. B-4A). The relationship holds within a country, for instance for accessions from Togo ($r^2 = 0.77$, $p < 0.0001$) and Senegal accessions ($r^2 = 0.78$, $p <$

0.0001) (Fig. B-4B). Phenotypic correlations, assessed based on Pearson's correlation coefficient to determine the relationship between adaptive traits (DFLo and PH) showed significant correlations among traits in each environment (Fig. B-5).

## Genome-wide association studies for flowering time and plant height

We assessed the effectiveness of the genome-wide SNP data for genetic dissection of complex quantitative traits using GWAS. For DFLo BLUPs, the GLM naive model identified many significant associations at Bonferroni correction 0.05 (Fig. 3-5A). A QTL was identified near *Ma6/ Ghd7* candidate gene between SNPs S6_651847 (top association in the region, ~45 kb away) and S6_699843 (within gene). A second QTL was identified between S9_54917833 and S9_54968379 at 43 kb and 4 kb from *SbCN8* candidate gene, respectively (Table B-3). The GLM showed an inflation of many significantly associated SNPs. The MLMM model identified eight QTL at Bonferroni threshold ($3.8 \times 10^{-7}$) (Fig. 5B). Some of these QTLs colocalized with known candidate genes, *Ma6* at 45 kb (S6_651847) and *SbCN8* at 381 kb (S9_55345348) (Table 3-3). The QTL near *Ma6*, S6_651847, was the top peak in the region in both GLM and MLMM and was at one gene away from *Ma6* (Fig. 3-5C and D). LD between the QTL, S6_651847 and SNPs near/within *Ma6* locus was moderate (Fig. 3-5E). After controlling for the population structure, the association of S6_651847 had an estimated effect size of 29 days and PVE of 25% (Table 3-3; Fig. 3-5F).

For PH, the GLM naive model identified several associations (Fig. 3-6A). The top association, S7_56232413 overlapped with the height QTL *qHT7.1* (Li *et al.*, 2015; Bouchet *et al.*, 2017). A second QTL was identified between SNPs S7_59955806 (top association in the region) and S7_59402662 located at 125 kb and 419 kb from the *Dw3 a priori* candidate gene, respectively (Table B-4). The MLMM identified three QTLs at the Bonferroni threshold (Fig. 3-6B; Table B-5). A putative SNP QTL, S7_59400476 was identified 421 kb away from *Dw3*, though below the Bonferroni threshold. After accounting for confounding population structure, the MLMM QTLs still had significant estimated effect sizes and contributed to high PVE for PH (Table B-5) after controlling for the population structure confounding effect. LD between S7_59955806 and S7_59400476 was high (though these two SNPs are separated by 555 kb from each other) but weak between these SNP QTLs and a variant in *Dw3* locus (Fig. 3-6E). Alleles at both SNP QTLs were associated with height differences of accessions across planting dates (Fig.

3-6F, G). The association of S7_59400476 with PH, which had the highest allelic effect estimate (73 cm) and PVE (41%), was confirmed using MLM ($p < 10^{-13}$) (Table B-5).

## Discussion

In the present study, we assembled 756 sorghum accessions from West African sorghum germplasm and characterized the genome-wide SNP variation of the panel. We demonstrated that this genome-wide SNP dataset is of sufficient quality for genomic and quantitative genetic analyses suitable for crop improvement through genomics-assisted breeding.

### A high-quality genomic resource

The strict data filtering criteria used before and after genotype imputation provided a final dataset with reduced number of SNPs (n = 159,101) many of which have impacts on protein-coding sequences (Fig. B-1). The quality control of the SNP dataset matched our expectations as the $F_{ST}$, PCA, and neighbor-joining analyses (Fig. 3-2, 4 and Fig. B-2) confirmed the expected structure of sorghum by botanical type and geographic region (Morris *et al.*, 2013; Lasky *et al.*, 2015; Bouchet *et al.*, 2017; Wang *et al.*, 2019). The validity of the SNP dataset was further confirmed based on GWAS with the identification of QTLs (Fig. 3-5, 6) that colocalized with known candidate loci, *Ma6* and *SbCN8* (Yang *et al.*, 2014; Murphy *et al.*, 2014) for days to flowering and *qHT7.1* and *Dw3* (Multani *et al.*, 2003; Li *et al.*, 2011, 2015) for plant height. Although regional diversity panels would limit confounding effects due to population structure in GWAS, the strong population structure observed in the WASAP appeared to increase false positive associations in the GLM. However, using the stepwise regression in MLMM though helped to control the inflation of *p*-values observed in the GLM (Fig. 3-5, 6).

Resolution of GWAS depends on linkage disequilibrium (LD) decay across the genome (Slatkin, 2008; Korte & Farlow, 2013). LD decay range was generally short in the WASAP and in the GDP (Fig. 3-1C) compared to the long LD range of the GDP reported in previous studies. In sorghum, studies have reported variation in LD decays from short (10–15 kb) to moderate (50–100 kb) (Hamblin *et al.*, 2005; Bouchet *et al.*, 2012) to about 150 kb in some studies with higher density of markers and larger population size (Morris *et al.*, 2013; Mace *et al.*, 2013b). The shorter LD range observed in the GDP in this study could be explained by the exclusion of North American breeding lines that share many common haplotypes (Klein *et al.*, 2008). Within the WASAP, the longer LD range in Togo accessions and shorter LD in Niger accessions (Fig. 3-

1C) are consistent with the limited number of genetic subpopulations observed in Togo compared to the strong genetic structure in Niger (Fig. 3-3C).

## Insights into hierarchical population structure in the West African sorghum

Sorghum genetic studies have identified population structure by botanical type and geographic region at regional scale (Deu *et al.*, 2006; Bouchet *et al.*, 2012; Morris *et al.*, 2013; Wang *et al.*, 2019) and at country level in the Senegal and Niger germplasm (Deu *et al.*, 2008; Faye *et al.*, 2019). The genetic diversity of the WASAP was structured by botanical type within each country of origin (Fig. 3 and 4). This finding is congruent with the $F_{ST}$ analysis (Table B-1) where botanical type and country of origin contributed to high and moderate genetic differentiation, respectively. The guinea type in the WASAP was split into three major subgroups (Fig. 3-4). One group was formed by Senegal and Mali accessions, a second group formed by Togo accessions (which clustered with Nigeria accessions), and a third group that was more related to durra and durra-caudatum types, formed predominantly by Senegal accessions. This third group (G-IV) clustered with guinea margaritiferum accessions from Niger and was more related to bicolor and wild sorghums in the GDP. Four groups of guinea were found in (Deu *et al.*, 2006), which included guinea from South Africa. Results did not show genetic differences between WASAP and WAS-GRIN populations from the same geographic origin. In contrast, genetic differences were observed between West African accessions and GDP where few subpopulations were formed almost entirely by West African sorghum accessions (e.g., G-VI and G-VII, Fig. 3-4).

The high genetic diversity observed within each of the four countries of the WASAP (Fig. 3-3C) is relevant for breeding programs in the region. Within the WASAP, all eight ancestral populations were found in the germplasm of each country, except in the Togo germplasm, which appeared to be less diverse. Altogether, the genetic diversity of WA sorghum germplasm is hierarchically structured by botanical type and subpopulation within botanical type, with many subpopulations distributed across countries.

## Suitability of genomic resources for GWAS

To establish the utility of our genome-wide SNP dataset for GWAS, we carried out GWAS for flowering time and plant height and demonstrated colocalization of QTL with known genes from previous studies. While flowering time genes and natural variants have been characterized in the US sorghum (Murphy *et al.*, 2014; Casto *et al.*, 2019), the genetic basis of

the substantial photoperiodic flowering time variation in West African sorghum is not yet known (Bhosale *et al.*, 2012). The QTL S6_651847 near *Ma6* (*Gdh7*) (Murphy *et al.*, 2014) highly contributed to the proportion of phenotypic variation of flowering time. This QTL was mapped in both GLM and MLMM (Fig. 3-5C, D), indicating that *Ma6* might be a major flowering time gene in the WA sorghum germplasm. Only two SNPs, which were in low LD with the QTL S6_651847 (Fig. 3-5E), were found within *Ma6* locus. Several of the other identified QTLs overlapped with flowering time QTLs found in other studies based on the sorghum QTL Atlas (Mace *et al.*, 2019) (Table B-6**)**. Our findings suggest a substantial oligogenic component conditioning flowering time in WA sorghums, indicating that photoperiodic flowering can be selected using markers from large effect QTLs. The two MLMM height QTLs (S5_30001948, and S9_38942669) accounted for 13.3% and 20.9%, respectively of height variation (Table B-5) but were not colocalized (within 500 kb) with known height genes. The identified QTLs with a major effect from this study would be cross validated using multi-parent mapping population and multi-year phenotypic data.

### Implications for sorghum improvement

This study demonstrates that breeding populations in each of the four countries in West Africa harbor sufficient genetic diversity. The hierarchical population structure observed in the WASAP at country level suggests the existence of multiple ancestral populations within the country but similar across West Africa. The increased kinship within each subpopulation enables the implementation of genomic selection within individual subpopulation and/or a subpopulation across breeding programs. Otherwise, the strong population structure could lead to biased prediction accuracy as a result from allele frequency differences among subpopulations (Isidro *et al.*, 2015). The SNP dataset could be used in genomic selection, genetic diversity, haplotype analyses, and GWAS analyses by the genetics community for global sorghum breeding. West African sorghum germplasm has been useful for global sorghum breeding, including durra and caudatum accessions that were sources of yellow endosperm and drought tolerance for US breeding programs (Rosenow & Dahlberg, 2000). Whole genome-resequencing would complement the GBS-SNP dataset to more easily identify causal variants for adaptive traits in the West African germplasm (Bellis *et al.*, 2020). The development of diagnostic markers for use in breeding will allow developing more rapidly locally adapted sorghum varieties through marker-assisted selection.

## Data Availability

The Supplemental Data S1 and SNP and phenotypic datasets generated from this study are available in the Dryad Data Repository under accession *https://doi.org/10.5061/dryad.k0p2ngf67*.

## Acknowledgements

This study is made possible by the support of the American People provided to the Feed the Future Innovation Lab for Collaborative Research on Sorghum and Millet through the United States Agency for International Development (USAID) under Cooperative Agreement No. AID-OAA-A-13-00047. The contents are the sole responsibility of the authors and do not necessarily reflect the views of USAID or the United States Government. The study was conducted using resources at the Integrated Genomics Facility and Beocat High-Performance Computing Cluster at Kansas State University.

# References


Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* 19: 1655–1664.

Bellis, E. S., Kelly, E. A., Lorts, C. M., Gao, H., DeLeo, V. L., Rouhan, G., Budden, A., Bhaskara, G. B., Hu, Z., Muscarella, R., Timko, M. P., Nebie, B., Runo, S. M., Chilcoat, N. D., Juenger, T. E., Morris, G. P., dePamphilis, C. W., & Lasky, J. R. (2020). Genomics of sorghum local adaptation to a parasitic plant. *Proceedings of the National Academy of Sciences*.

Bhosale, S. U., Stich, B., Rattunde, H. F. W., Weltzien, E., Haussmann, B. I., Hash, C. T., Ramu, P., Cuevas, H. E., Paterson, A. H., Melchinger, A. E., & Parzies, H. K. (2012). Association analysis of photoperiodic flowering time genes in west and central African sorghum [Sorghum bicolor (L.) Moench]. *BMC Plant Biology*, *12*, 32.

Bouchet S, Olatoye MO, Marla SR, Perumal R, Tesso T, Yu J, Tuinstra M, Morris GP. 2017. Increased Power To Dissect Adaptive Traits in Global Sorghum Diversity Using a Nested Association Mapping Population. *Genetics* 206: 573–585.

Bouchet S, Pot D, Deu M, Rami J-F, Billot C, Perrier X, Rivallan R, Gardes L, Xia L, Wenzl P, *et al.*2012. Genetic Structure, Linkage Disequilibrium and Signature of Selection in Sorghum: Lessons from Physically Anchored DArT Markers. *PLOS ONE* 7: e33470.

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. 2007. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23: 2633–2635.

Browning BL, Browning SR. 2016. Genotype Imputation with Millions of Reference Samples. *The American Journal of Human Genetics* 98: 116–126.

Cao K, Zhou Z, Wang Q, Guo J, Zhao P, Zhu G, Fang W, Chen C, Wang X, Wang X, *et al.*2016. Genome-wide association study of 12 agronomic traits in peach. *Nature Communications* 7: 13246.

Casto, A. L., Mattison, A. J., Olson, S. N., Thakran, M., Rooney, W. L., & Mullet, J. E. (2019). Maturity2, a novel regulator of flowering time in Sorghum bicolor, increases expression of SbPRR37 and SbCO in long days delaying flowering. *PLOS ONE*, *14*(4), e0212154.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* 6: 80–92.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, *et al.*2011. The variant call format and VCFtools. *Bioinformatics* 27: 2156–2158.

Deu M, Gonzalez-de-Leon D, Glaszmann J-C, Degremont I, Chantereau J, Lanaud C, Hamon P. 1994. RFLP diversity in cultivated sorghum in relation to racial differentiation. *Theoretical and Applied Genetics* 88: 838–844.

Deu M, Rattunde F, Chantereau J. 2006. A global view of genetic diversity in cultivated sorghums using a core collection. *Genome* 49: 168–180.

Deu M, Sagnard F, Chantereau J, Calatayud C, Hérault D, Mariac C, Pham J-L, Vigouroux Y, Kapran I, Traore PS, *et al.*2008. Niger-wide assessment of in situ sorghum genetic diversity with microsatellite markers. *Theoretical and Applied Genetics* 116: 903–913.

Doggett H. 1988. *Sorghum*. Longman Scientific & Technical.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species (L Orban, Ed.). *PLoS ONE* 6: e19379.

Faye JM, Maina F, Hu Z, Fonceka D, Cisse N, Morris GP. 2019. Genomic signatures of adaptation to Sahelian and Soudanian climates in sorghum landraces of Senegal. *Ecology and Evolution* 9: 6038–6051.

Felderhoff TJ, Murray SC, Klein PE, Sharma A, Hamblin MT, Kresovich S, Vermerris W, Rooney WL. 2012. QTLs for Energy-related Traits in a Sweet × Grain Sorghum [ Sorghum bicolor (L.) Moench] Mapping Population. *Crop Science* 52: 2040–2049.

Feltus FA, Hart GE, Schertz KF, Casa AM, Kresovich S, Abraham S, Klein PE, Brown PJ, Paterson AH. 2006. Alignment of genetic maps and QTLs between inter- and intra-specific sorghum populations. *Theoretical and Applied Genetics* 112: 1295.

Foley JA, Ramankutty N, Brauman KA, Cassidy ES, Gerber JS, Johnston M, Mueller ND, O'Connell C, Ray DK, West PC, *et al.*2011. Solutions for a cultivated planet. *Nature* 478: 337–342.

Folkertsma RT, Rattunde HFW, Chandra S, Raju GS, Hash CT. 2005. The pattern of genetic diversity of Guinea-race Sorghum bicolor (L.) Moench landraces as revealed with SSR markers. *Theoretical and Applied Genetics* 111: 399–409.

Fu Y-B. 2017. The Vulnerability of Plant Genetic Resources Conserved Ex Situ. *Crop Science* 57: 2314–2328.

Gapare W, Conaty W, Zhu Q-H, Liu S, Stiller W, Llewellyn D, Wilson I. 2017. Genome-wide association study of yield components and fibre quality traits in a cotton germplasm diversity panel. *Euphytica* 213: 66.

Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES. 2014. TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. *PLOS ONE* 9:

e90346.

Hamblin MT, Fernandez MGS, Casa AM, Mitchell SE, Paterson AH, Kresovich S. 2005. Equilibrium Processes Cannot Explain High Levels of Short- and Medium-Range Linkage Disequilibrium in the Domesticated Grass Sorghum bicolor. *Genetics* 171: 1247–1256.

Hammer K, Teklu Y. 2008. Plant Genetic Resources: Selected Issues from Genetic Erosion to Genetic Engineering. *Journal of Agriculture and Rural Development in the Tropics and Subtropics* Volume 109: 15–50.

Hernandez, R.D., L.H. Uricchio, K. Hartman, C. Ye, A. Dahl, and N. Zaitlen. 2019. Ultrarare variants drive substantial cis heritability of human gene expression. Nature Genetics 51:1349–1355. doi:10.1038/s41588-019-0487-7.

Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z, *et al.*2010. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature Genetics* 42: 961–967.

Isidro J, Jannink J-L, Akdemir D, Poland J, Heslot N, Sorrells ME. 2015. Training set optimization under population structure in genomic selection. *Theoretical and Applied Genetics* 128: 145–158.

ISRA. 2005. Bilan de la recherche agricole et agroalimentaire au Sénégal. *Institut sénégalais de recherches agricoles*: 524.

Jordan DR, Mace ES, Cruickshank AW, Hunt CH, Henzell RG. 2011. Exploring and Exploiting Genetic Variation from Unadapted Sorghum Germplasm in a Breeding Program. *Crop Science* 51: 1444–1457.

Klein RR, Mullet JE, Jordan DR, Miller FR, Rooney WL, Menz MA, Franks CD, Klein PE. 2008. The Effect of Tropical Sorghum Conversion and Inbred Development on Genome Diversity as Revealed by High-Resolution Genotyping. *Crop Science* 48: S-12-S-26.

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

Lasky JR, Upadhyaya HD, Ramu P, Deshpande S, Hash CT, Bonnette J, Juenger TE, Hyma K, Acharya C, Mitchell SE, *et al.*2015. Genome-environment associations in sorghum landraces predict adaptive traits. *Science Advances* 1: e1400218.

Leiser WL, Rattunde HFW, Weltzien E, Cisse N, Abdou M, Diallo A, Tourè AO, Magalhaes JV, Haussmann BI. 2014. Two in one sweep: aluminum tolerance and grain yield in P-limited soils are associated to the same genomic region in West African Sorghum. *BMC Plant Biology* 14.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754–1760.

Li J, Jiang J, Qian Q, Xu Y, Zhang C, Xiao J, Du C, Luo W, Zou G, Chen M, *et al.*2011. Mutation of Rice BC12/GDD1, Which Encodes a Kinesin-Like Protein That Binds to a GA Biosynthesis Gene Promoter, Leads to Dwarfism with Impaired Cell Elongation. *The Plant Cell* 23: 628–640.

Li X, Li X, Fridman E, Tesso TT, Yu J. 2015. Dissecting repulsion linkage in the dwarfing gene Dw3 region for sorghum plant height provides insights into heterosis. *Proceedings of the National Academy of Sciences of the United States of America* 112: 11823–11828.

Lin YR, Schertz KF, Paterson AH. 1995. Comparative analysis of QTLs affecting plant height and maturity across the Poaceae, in reference to an interspecific sorghum population. *Genetics* 141: 391–411.

Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang Z. 2012. GAPIT: genome association and prediction integrated tool. *Bioinformatics* 28: 2397–2399.

Mace ES, Hunt CH, Jordan DR. 2013a. Supermodels: sorghum and maize provide mutual insight into the genetics of flowering time. *Theoretical and Applied Genetics* 126: 1377–1395.

Mace E, Innes D, Hunt C, Wang X, Tao Y, Baxter J, Hassall M, Hathorn A, Jordan D. 2019. The Sorghum QTL Atlas: a powerful tool for trait dissection, comparative genomics and crop improvement. *Theoretical and Applied Genetics* 132: 751–766.

Mace ES, Tai S, Gilding EK, Li Y, Prentis PJ, Bian L, Campbell BC, Hu W, Innes DJ, Han X, *et al.*2013b. Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. *Nature Communications* 4: 2320.

Maina F, Bouchet S, Marla SR, Hu Z, Wang J, Mamadou A, Abdou M, Saïdou A-A, Morris GP. 2018. Population genomics of sorghum (Sorghum bicolor) across diverse agroclimatic zones of Niger. *Genome* 61: 223–232.

McCormick RF, Truong SK, Sreedasyam A, Jenkins J, Shu S, Sims D, Kennedy M, Amirebrahimi M, Weers BD, McKinley B, *et al.*2018. The Sorghum bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *The Plant Journal* 93: 338–354.

Meyer RS, Purugganan MD. 2013. Evolution of crop species: genetics of domestication and diversification. *Nature Reviews Genetics* 14: 840–852.

Mendiburu, F. (2009). *Agricolae: Statistical Procedures for Agricultural Research.* https://rdrr.io/cran/agricolae/man/agricolae-package.html

Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, Riera-Lizarazu O, Brown PJ, Acharya CB, Mitchell SE, *et al.*2013. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proceedings of the National Academy of Sciences of the United States of America* 110: 453–458.

Multani DS, Briggs SP, Chamberlin MA, Blakeslee JJ, Murphy AS, Johal GS. 2003. Loss of an MDR Transporter in Compact Stalks of Maize br2 and Sorghum dw3 Mutants. *Science* 302: 81–84.

Mundia CW, Secchi S, Akamani K, Wang G. 2019. A Regional Comparison of Factors Affecting Global Sorghum Production: The Case of North America, Asia and Africa's Sahel. *Sustainability* 11: 2135.

Murphy RL, Morishige DT, Brady JA, Rooney WL, Yang S, Klein PE, Mullet JE. 2014. Ghd7 ( Ma 6 ) Represses Sorghum Flowering in Long Days: Ghd7 Alleles Enhance Biomass Accumulation and Grain Production. *The Plant Genome* 7.

Olatoye MO, Hu Z, Maina F, Morris GP. 2018. Genomic Signatures of Adaptation to a Precipitation Gradient in Nigerian Sorghum. *G3: Genes, Genomes, Genetics*: g3.200551.2018.

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics (Oxford, England)* 20: 289–290.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, *et al.*2009. The Sorghum bicolor genome and the diversification of grasses. *Nature* 457: 551–556.

Peloso GM, Rader DJ, Gabriel S, Kathiresan S, Daly MJ, Neale BM. 2016. Phenotypic extremes in rare variant study designs. *European Journal of Human Genetics* 24: 924–930.

Peterson, B. G., Carl, P., Boudt, K., Bennett, R., Ulrich, J., Eric Zivot, Lestel, M., Balkissoon, K., & Wuertz, D. (2014). *PerformanceAnalytics: Econometric Tools for Performance and Risk Analysis. Version 1.4.4000 from R-Forge*.

Poland JA, Brown PJ, Sorrells ME, Jannink J-L. 2012. Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. *PLOS ONE* 7: e32253.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, *et al.*2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* 81: 559–575.

R Core Team RC. 2016. *A language and environment for statistical computing. R Foundation for statistical computing, 2015; Vienna, Austria*.

Sangma BH. 2013. Genetic characterization of flowering time in sorghum - UQ eSpace.

Sattler FT, Sanogo MD, Kassari IA, Angarawai II, Gwadi KW, Dodo H, Haussmann BIG. 2018. Characterization of West and Central African accessions from a pearl millet reference collection for agro-morphological traits and *Striga* resistance. *Plant Genetic Resources: Characterization and Utilization* 16: 260–272.

Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, Long Q, Nordborg M. 2012. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genetics* 44: 825–830.

Shin J-H, Blay S, McNeney B, Graham J. 2006. LDheatmap: An R Function for Graphical Display of Pairwise Linkage Disequilibria Between Single Nucleotide Polymorphisms | Shin | Journal of Statistical Software. 16.

Slatkin M. 2008. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics* 9: 477–485.

Tilman D, Balzer C, Hill J, Befort BL. 2011. Global food demand and the sustainable intensification of agriculture. *Proceedings of the National Academy of Sciences* 108: 20260–20264.

Wang J, Hu Z, Upadhyaya HD, Morris GP. 2019. Genomic signatures of seed mass adaptation to global precipitation gradients in sorghum. *Heredity*: 1.

Wang X, Mace E, Hunt C, Cruickshank A, Henzell R, Parkes H, Jordan D. 2014. Two distinct classes of QTL determine rust resistance in sorghum. *BMC Plant Biology* 14: 366.

Yang S, Murphy RL, Morishige DT, Klein PE, Rooney WL, Mullet JE. 2014. Sorghum Phytochrome B Inhibits Flowering in Long Days by Activating Expression of SbPRR37 and SbGHD7, Repressors of SbEHD1, SbCN8 and SbCN12. *PLOS ONE* 9: e105352.

Yano K, Yamamoto E, Aya K, Takeuchi H, Lo P, Hu L, Yamasaki M, Yoshida S, Kitano H, Hirano K, *et al.*2016. Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nature Genetics* 48: 927–934.

Zhang C, Dong S-S, Xu J-Y, He W-M, Yang T-L. 2019. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* 35: 1786–1788.

Zhang D, Kong W, Robertson J, Goff VH, Epps E, Kerr A, Mills G, Cromwell J, Lugin Y, Phillips C, *et al.*2015. Genetic analysis of inflorescence and plant height components in sorghum (Panicoidae) and comparative genetics with rice (Oryzoidae). *BMC Plant Biology* 15: 107.

Zhao Y, Qiang C, Wang X, Chen Y, Deng J, Jiang C, Sun X, Chen H, Li J, Piao W, *et al.*2019. New alleles for chlorophyll content and stay-green traits revealed by a genome wide association study in rice ( Oryza sativa ). *Scientific Reports* 9: 2541.

Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 28: 3326–3328.

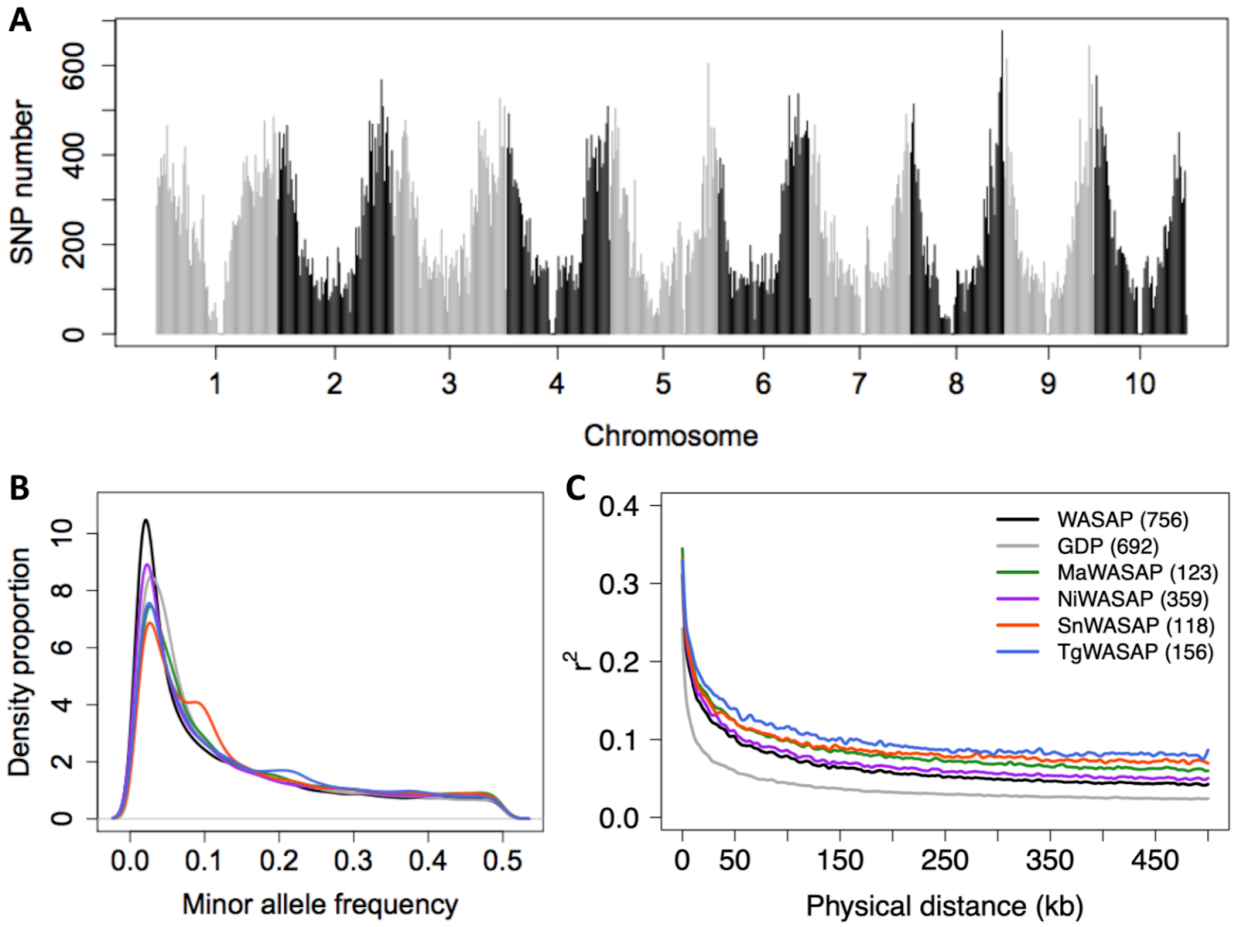



**Figure 3-1. Genome-wide SNP variation in the WASAP and GDP.** (A) Distribution of the SNP data across the 10 sorghum chromosomes in the WASAP. Minor allele frequency distribution (B) and Linkage disequilibrium decay along the genome (C) of the SNP data in the whole WASAP, within country in WASAP–Mali (MaWASAP), Niger (NiWASAP), Senegal (SnWASAP) and Togo (TgWASAP), and in the global sorghum diversity panel (GDP).

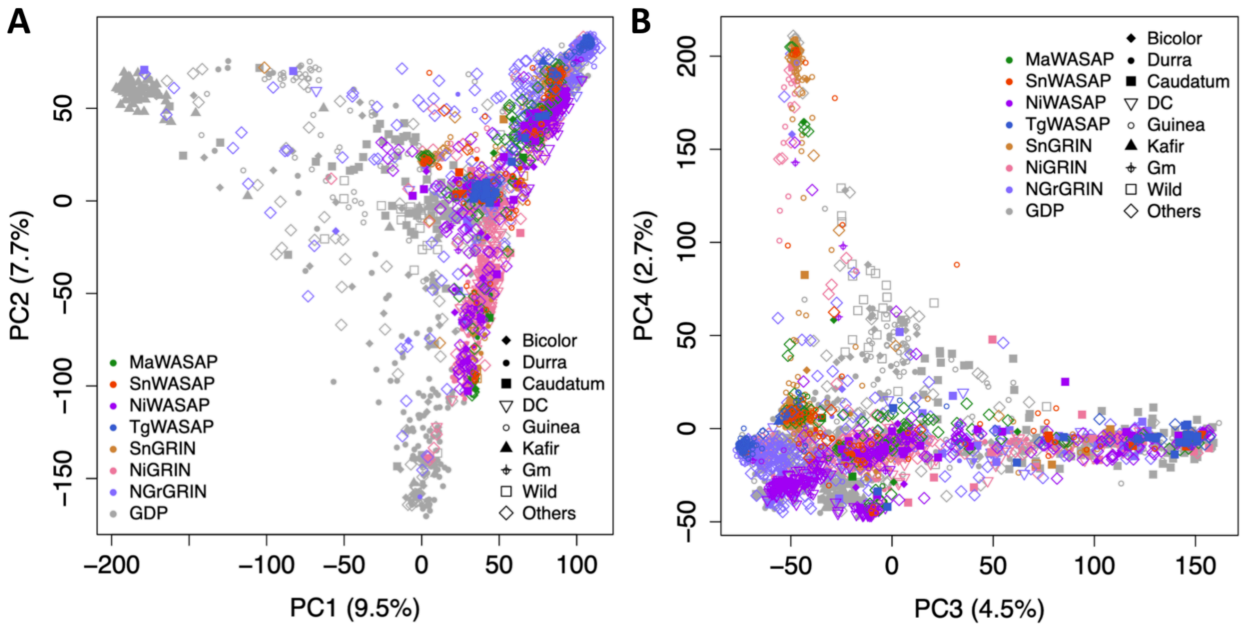



**Figure 3-2. Principal component analysis of genome-wide SNP variation.** Scatterplots of the first and second axes (A) and the third and fourth axes (B) of genome-wide SNP variation in the WASAP in relationship with other West African sorghums in GRIN and global sorghum diversity panel. The color-codes indicate country of origin for WASAP accessions (MaWASAP, Mali; NiWASAP, Niger; SnWASAP, Senegal and TgWASAP, Togo), the West African accessions in GRIN (SnGRIN, Senegal, Gambia and Mauritania; NiGRIN, Niger; NGrGRIN, Nigeria), and the global sorghum diversity panel (GDP). The symbols indicate botanical types where DC and Gm correspond to durra-caudatum intermediate and guinea-margaritiferum types, respectively.

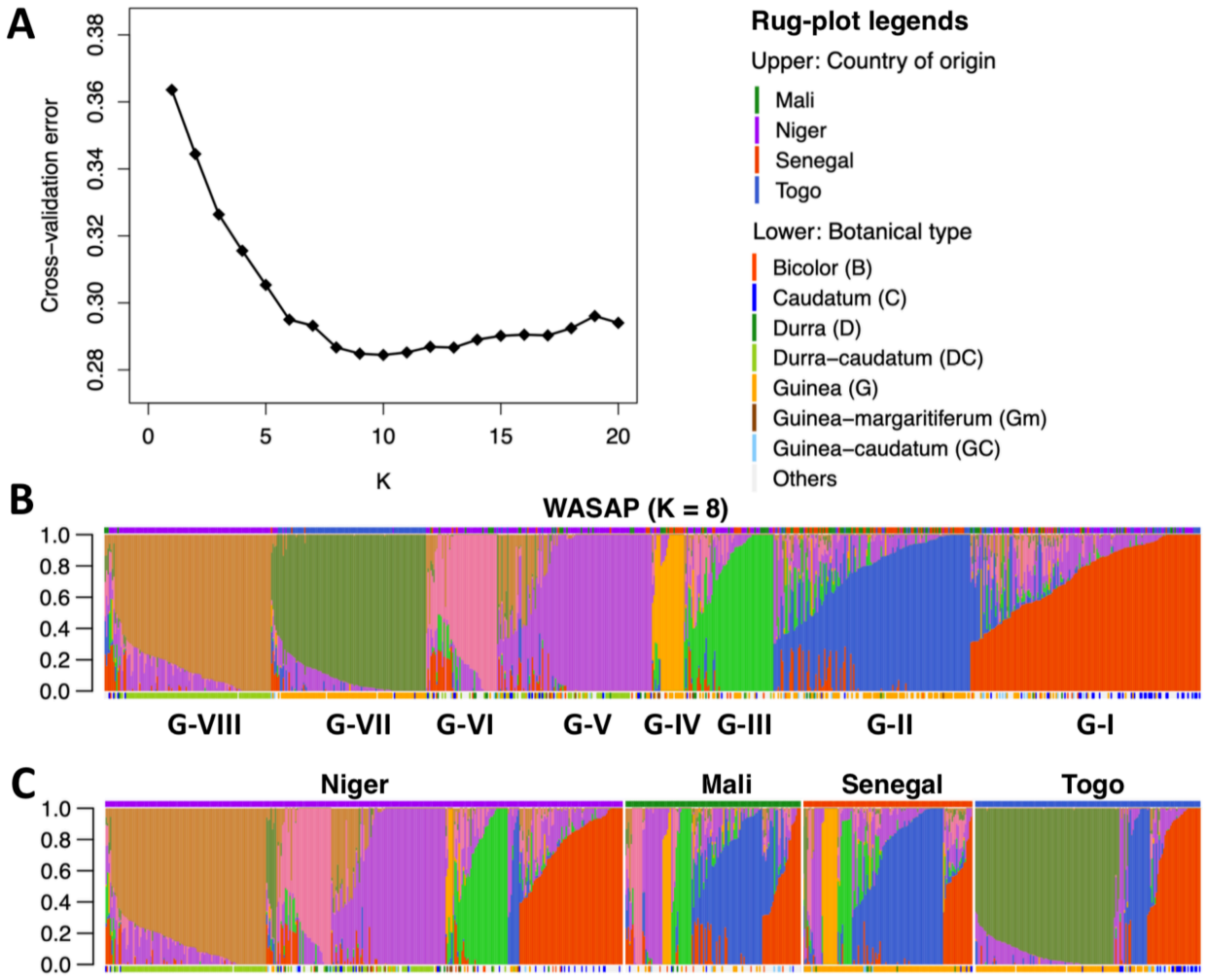



**Figure 3-3. Genetic ancestry analysis of the WASAP.** (A) Five-fold cross-validation error from the ADMIXTURE model using 60,749 SNPs for K = 2–20. Ancestral genetic groups of the WASAP at K = 8 ancestral populations (B) ordered by ancestry fraction and (C) ordered by country then by ancestry fraction. Each vertical bar plot on the x-axis represents ancestry fraction from the eight ancestral populations (G-I to G-VIII) indicated with a different arbitrary color for each accession. Upper rug-plots indicate countries of origin. Lower rug-plots indicate botanical type ("Others" include rare intermediate types and accessions of unknown botanical type). Ancestry fractions for each accession are available in Supplemental Data S1.

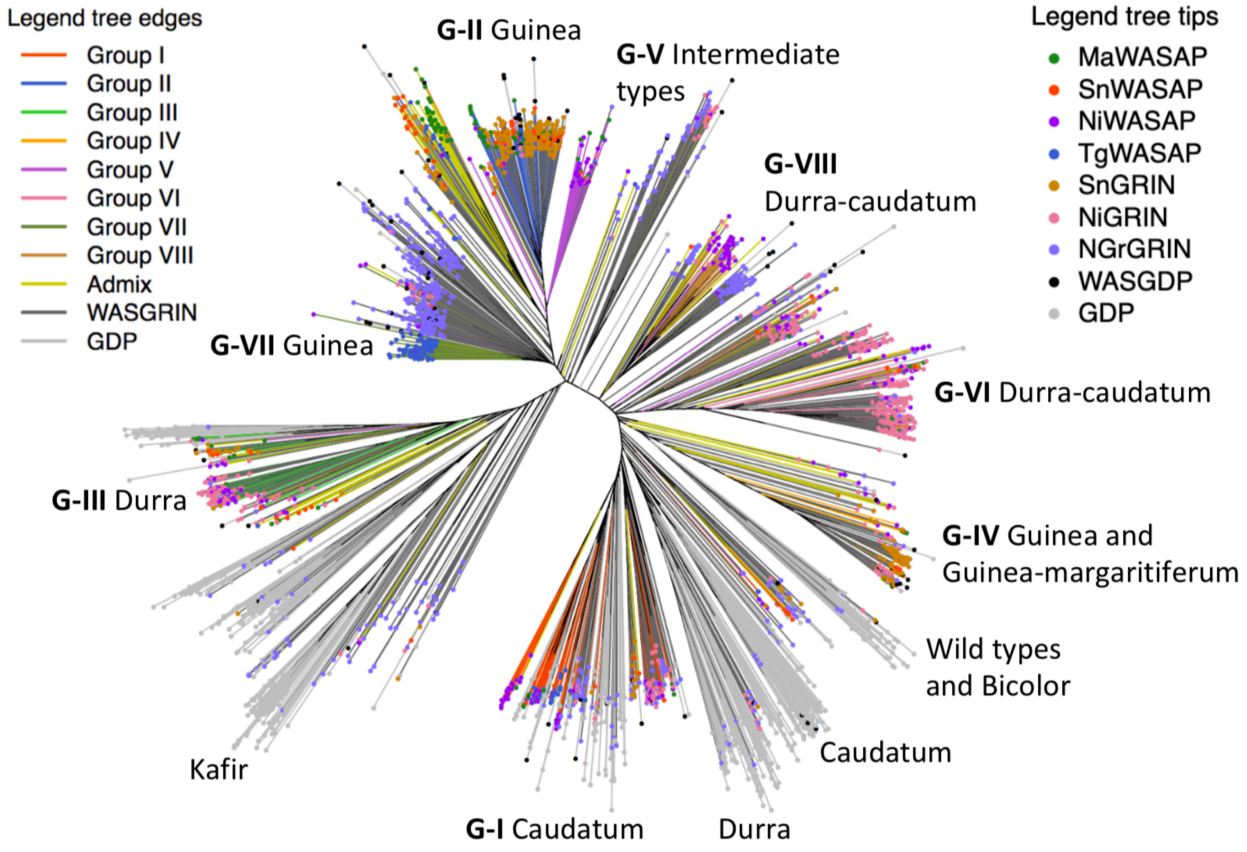



**Figure 3-4. Neighbor-joining analysis of the WASAP.** Clustering of the WASAP accessions (MaWASAP, Mali; NiWASAP, Niger; SnWASAP, Senegal and TgWASAP, Togo) in relationship with other West African sorghums in GRIN (SnGRIN, Senegal, Gambia and Mauritania; NiGRIN, Niger; NGrGRIN, Nigeria) and global sorghum diversity panel (GDP). The color-coding of the tree edges is based on the ADMIXTURE ancestral populations (G-I to G-VIII, including admixed accessions) of the WASAP. The edges in yellow, dark gray, and light gray represent admixed WASAP accessions (< 0.6 ancestry fraction), WASGRIN accessions, and GDP accessions, respectively. The color-coding of the tree tips indicate accessions origin, with black tips indicating West African sorghum accessions in the GDP (WASGDP).

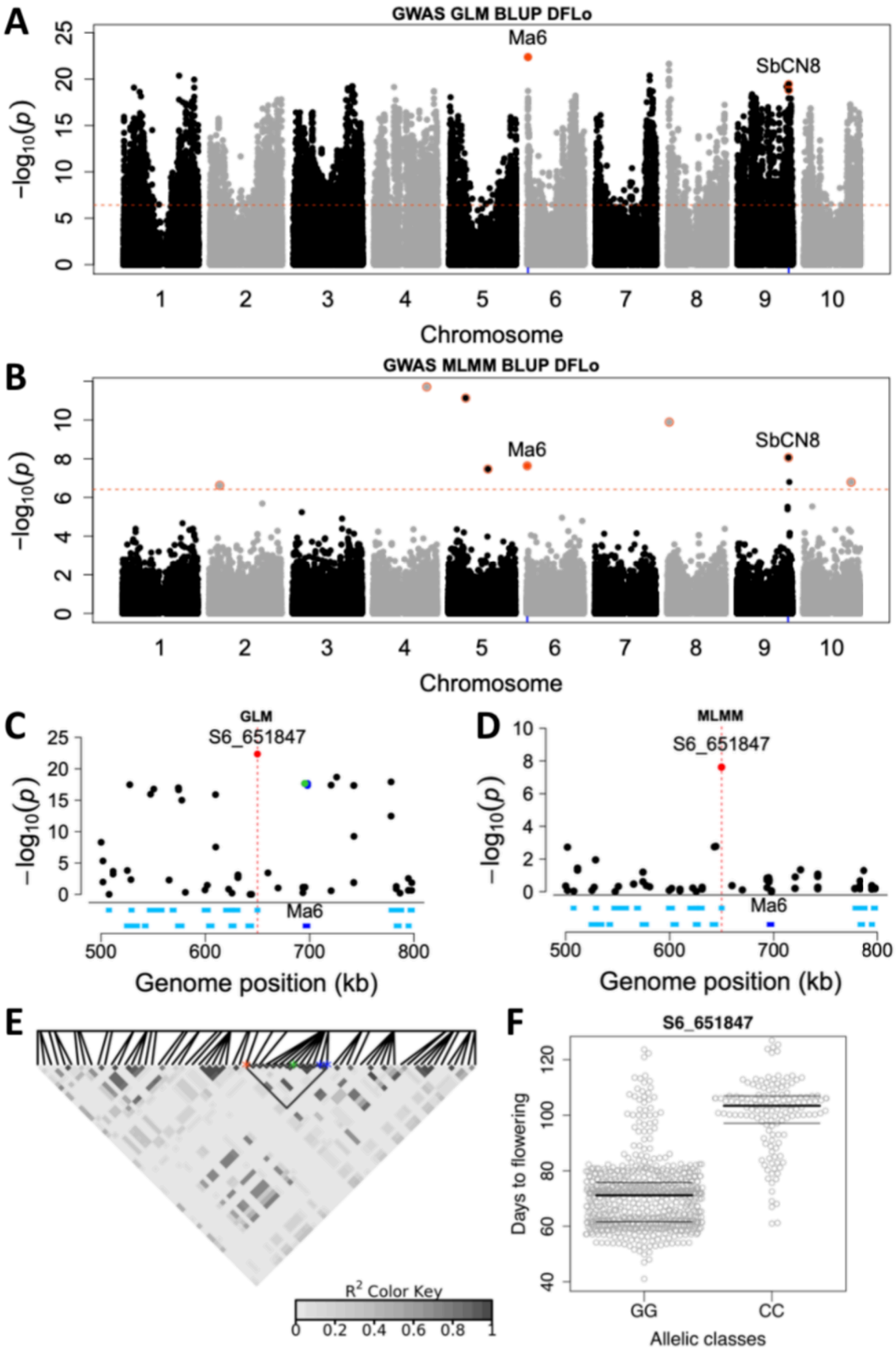

A
GWAS GLM BLUP DFLo
Ma6
SbCN8
−log10(p)
Chromosome
B
GWAS MLMM BLUP DFLo
Ma6
SbCN8
−log10(p)
Chromosome
C
GLM
S6_651847
Ma6
−log10(p)
Genome position (kb)
D
MLMM
S6_651847
Ma6
−log10(p)
Genome position (kb)
E
R2 Color Key
F
S6_651847
Days to flowering
GG
CC
Allelic classes

**Figure 3-5. GWAS for days to flowering (DFLo) under rainfed conditions.** Manhattan plots of DFLo based on (A) the GLM and (B) the MLMM. The horizontal red dashed line represents the Bonferroni significance threshold at 0.05. The rug plots indicate the position of colocalizing candidate genes, *Ma6* and *SbCN8* with QTLs. Regional Manhattan plot of a 150 kb region on chromosome 6 around the QTL S6_651847 that colocalized with *Ma6* from (C) GLM and (D) MLMM. The green and blue peaks are SNP QTLs at 160 bp from and within *Ma6*, respectively. The dark blue segment indicates the genomic position of *Ma6*. (E) LD heatmap of a 150 kb region surrounding the QTL S6_651847. The red, green, and blue asterisks indicate the S6_651847, SNPs at 160 bp from *Ma6* and within *Ma6*, respectively. (F) Days to flowering across planting dates by allelic classes of the QTL S6_651847.

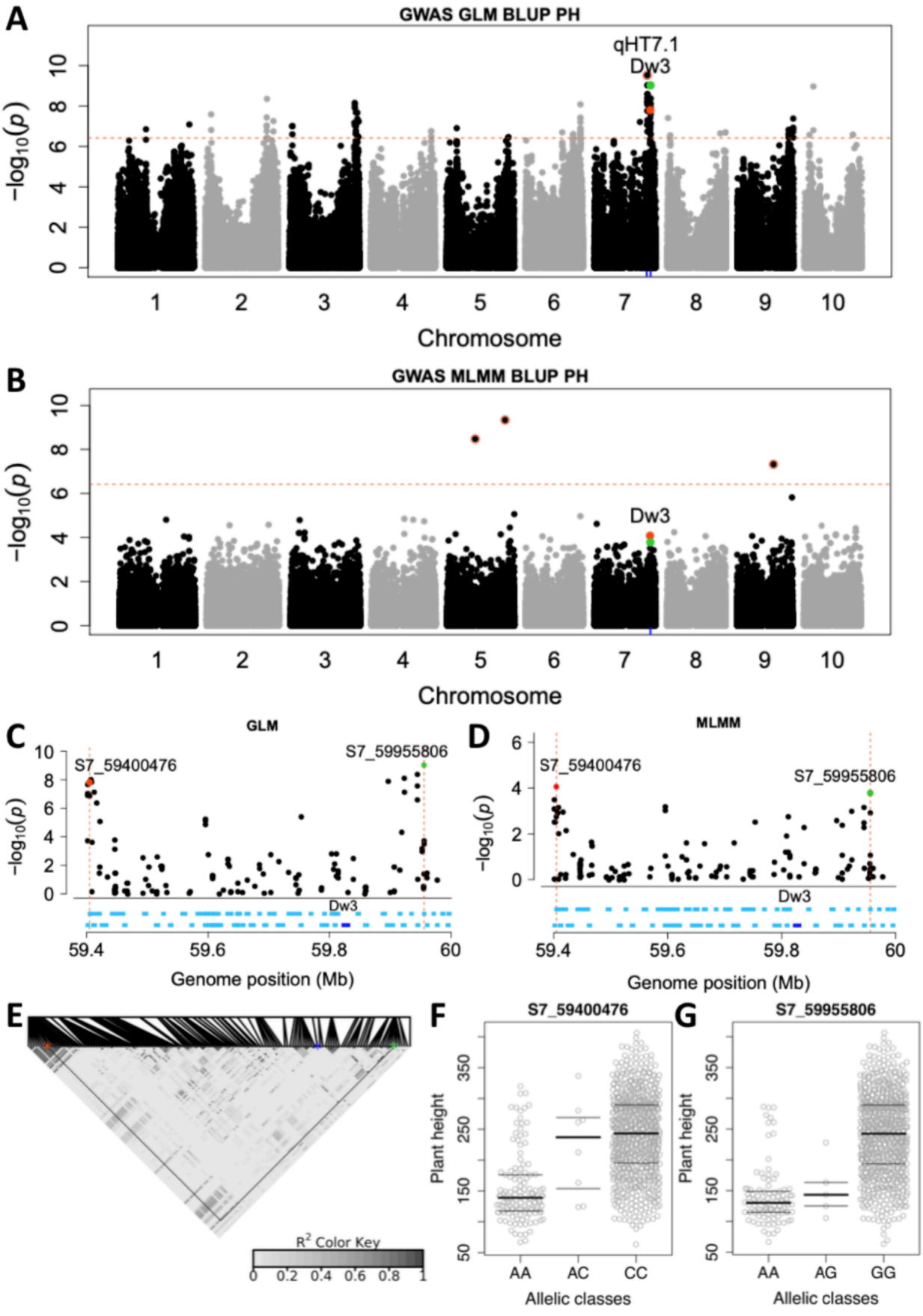

A
GWAS GLM BLUP PH
qHT7.1
Dw3
−log10(p)
Chromosome
B
GWAS MLMM BLUP PH
Dw3
−log10(p)
Chromosome
C
GLM
S7_59400476
S7_59955806
Dw3
−log10(p)
Genome position (Mb)
D
MLMM
S7_59400476
S7_59955806
Dw3
−log10(p)
Genome position (Mb)
E
R2 Color Key
F
S7_59400476
Plant height
AA
AC
CC
Allelic classes
G
S7_59955806
Plant height
AA
AG
GG
Allelic classes

**Figure 3-6. GWAS for plant height (PH) under rainfed conditions.** Manhattan plots of PH based on (A) the GLM and (B) the MLMM. The horizontal dashed line represents the Bonferroni significance threshold at 0.05. Rug plots on chromosome 7 indicate the position of the candidate gene, *Dw3* and *qPH7.1*. Regional Manhattan plot of a 600 kb region on chromosome 7 surrounding the QTL between S7_59400476 and S7_59955806 that colocalizes with *Dw3* from (C) GLM and (D) MLMM. The red and green peaks are top SNPs in MLMM and GLM, respectively. The dark blue segment indicates the genomic position of *Dw3*. (E) LD heatmap of genomic region between SNPs S7_59400476 and S7_59955806. The red, green, and blue asterisks indicate the S7_59400476, S7_59955806, and a SNP within *Dw3*, respectively. Days to flowering across planting dates by allelic classes of SNP QTLs (F) S7_59400476 and (G) S7_59955806.

**Table 3-1. Number of accessions in each sorghum collection/panel used in this study.**

| | Collection/panel | | |
|---|---|---|---|
| **Country of origin** | **WASAP** | **WAS-GRIN** | **GDP** |
| Mali | 123 | - | 15 |
| Niger | 359 | 515 | 12 |
| Senegal | 118 | 346 | 8 |
| Togo | 156 | - | 3 |
| Gambia | - | 60 | 4 |
| Mauritania | - | 15 | - |
| Nigeria | - | 607 | 38 |
| Other West Africa | - | - | 18 |
| *Total from WA* | 756 | 1543 | 98 |
| Central Africa | | | 44 |
| North Africa | | | 10 |
| South Africa | | | 146 |
| East Africa | | | 223 |
| Middle East | | | 37 |
| East Asia | | | 26 |
| South Asia | | | 99 |
| Other | | | 9 |
| *Total* | 756 | 1543 | 692 |

WASAP, West African sorghum association panel; WAS-GRIN, West African sorghum in USDA-GRIN; WA, West Africa; GDP, global sorghum diversity panel.

**Table 3-2. Descriptive statistics and phenotypic variation across early (Hiv1) and late (Hiv2) planting date experiments under rainfed conditions.**

| Trait | Range | Mean ± SD | CV (%) | G (%) | E (%) | GxE (%) | $H^2$ |
|---|---|---|---|---|---|---|---|
| DFLo | 40–128 | 78 ± 15 | 21 | $79^{***}$ | $1^{ns}$ | $9^{***}$ | 0.88 |
| PH (cm) | 65–410 | 214 ± 75 | 35 | $57^{***}$ | $3^{*}$ | $20^{***}$ | 0.74 |

DFLo, days to flowering; PH, plant height; SD, Standard deviation; CV, coefficient of variation; G, genotype and E, environment variances; $H^2$, broad-sense heritability; Significance levels of variance components for superscripts, ***$p < 0.001$ and *$p < 0.05$; ns, not significant.

**Table 3-3. Quantitative-trait loci associated with days to flowering BLUPs across early and late planting date experiments using the MLMM.**

| QTL SNP[a] | MLMM *p*-value | MAF | Effect size | PVE[b] (%) | MLM *p*-value | Distance to locus (kb) | Locus name |
|---|---|---|---|---|---|---|---|
| S4_57407080 | $<10^{-12}$ | 0.02 | 10 | 7 | 0.002 | | |
| S5_18513685 | $<10^{-12}$ | 0.10 | 12 | 9 | 0.0008 | | |
| S8_2206437 | $<10^{-10}$ | 0.04 | 28 | 11 | 0.0001 | | |
| S9_55345348 | $<10^{-9}$ | 0.04 | 15 | 9 | 0.0005 | 381 | *SbCN8* |
| S6_651847 | $<10^{-8}$ | 0.25 | 29 | 25 | $<10^{-11}$ | 45 | *Ma6* |
| S5_42404204 | $<10^{-8}$ | 0.12 | 6 | 16 | 0.0001 | | |
| S10_51083132 | $<10^{-7}$ | 0.28 | 7 | 5 | 0.03 | | |
| S2_11509143 | $<10^{-7}$ | 0.04 | 5 | 0.3 | 0.6 | | |

[a] Digits before and after underscore indicate chromosome number and SNP position on the genome, respectively; [b] ADMIXTURE ancestry fractions at K = 8 were used as fixed effect covariates; BLUP, best linear unbiased prediction; MLMM, multi-locus mixed-linear model; QTL, quantitative-trait loci; MAF, minor allele frequency; GLM, general linear model; PVE, proportion of variance explained; MLM, mixed-linear model.

# Supplemental Materials Chapter 3

This section includes the supplemental figures and tables for the chapter 3

A Genomic Resource for Genetics, Physiology, and Breeding of West African Sorghum.

| Type (alphabetical order) | Count | Percent |
|---|---|---|
| DOWNSTREAM | 60,102 | 22.146% |
| EXON | 14,186 | 5.227% |
| INTERGENIC | 116,339 | 42.868% |
| INTRON | 17,382 | 6.405% |
| SPLICE_SITE_ACCEPTOR | 86 | 0.032% |
| SPLICE_SITE_DONOR | 95 | 0.035% |
| SPLICE_SITE_REGION | 642 | 0.237% |
| UPSTREAM | 59,917 | 22.078% |
| UTR_3_PRIME | 1,820 | 0.671% |
| UTR_5_PRIME | 817 | 0.301% |

| Type (alphabetical order) | Count | Percent |
|---|---|---|
| HIGH | 619 | 0.228% |
| LOW | 6,814 | 2.511% |
| MODERATE | 7,654 | 2.82% |
| MODIFIER | 256,299 | 94.441% |

| Type (alphabetical order) | Count | Percent |
|---|---|---|
| MISSENSE | 7,689 | 71.84% |
| NONSENSE | 411 | 3.84% |
| SILENT | 2,603 | 24.32% |

**Figure B- 1. Genome-wide single nucleotide polymorphisms in the WASAP.** SNP annotation effect of 159,101 SNPs with MAF > 0.01.

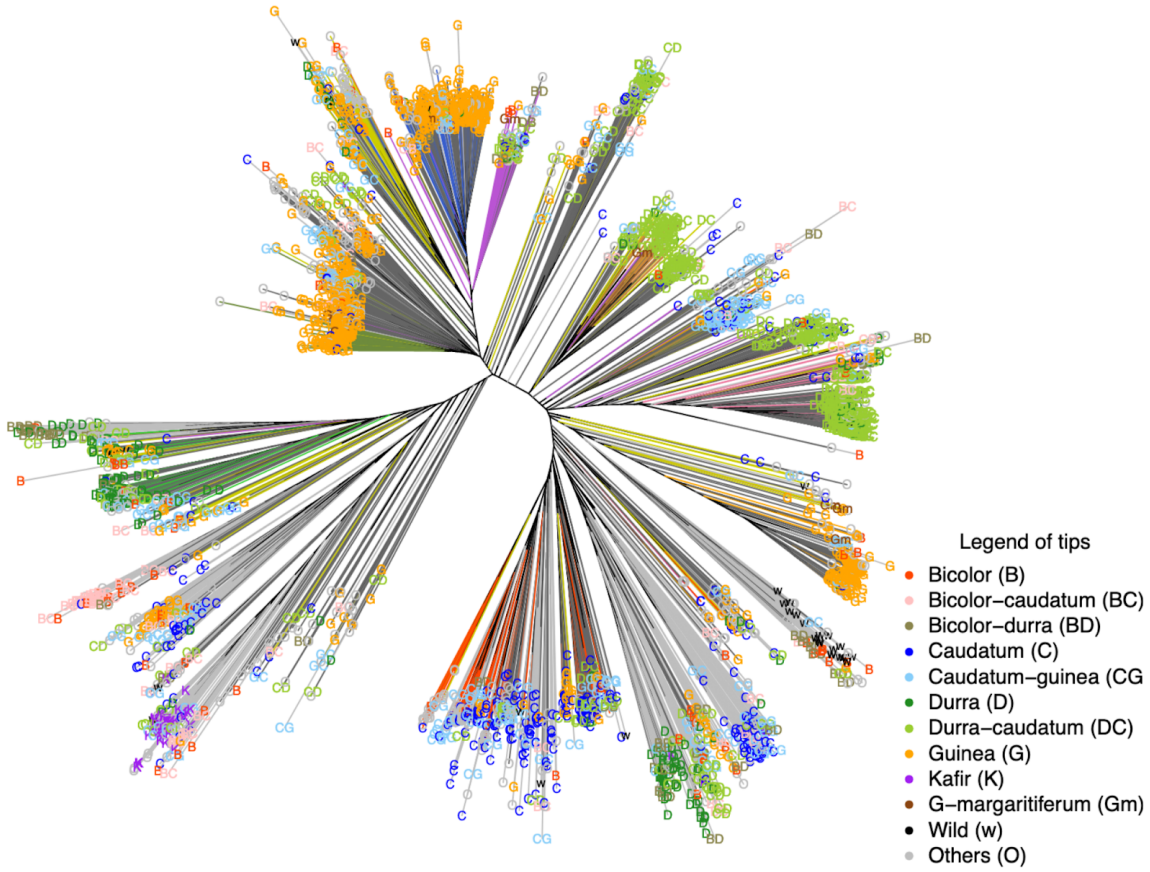



**Figure B- 2. Unrooted Neighbor-joining tree of WASAP ancestral populations in relationship with other West African sorghums in GRIN (WAS-GRIN) and the global sorghum diversity panel (GDP).** The tree edges and tips are color-coded based on the ADMIXTURE ancestral populations and the accessions origin, respectively. Tree edges in yellow, darkgray, and gray represent admixed (< 0.6 ancestry fraction) accessions in the WASAP, WAS-GRIN (SnGRIN, Senegal, Gambia and Mauritania; NiGRIN, Niger; NGrGRIN, Nigeria), and global sorghum diversity panel (GDP), respectively.

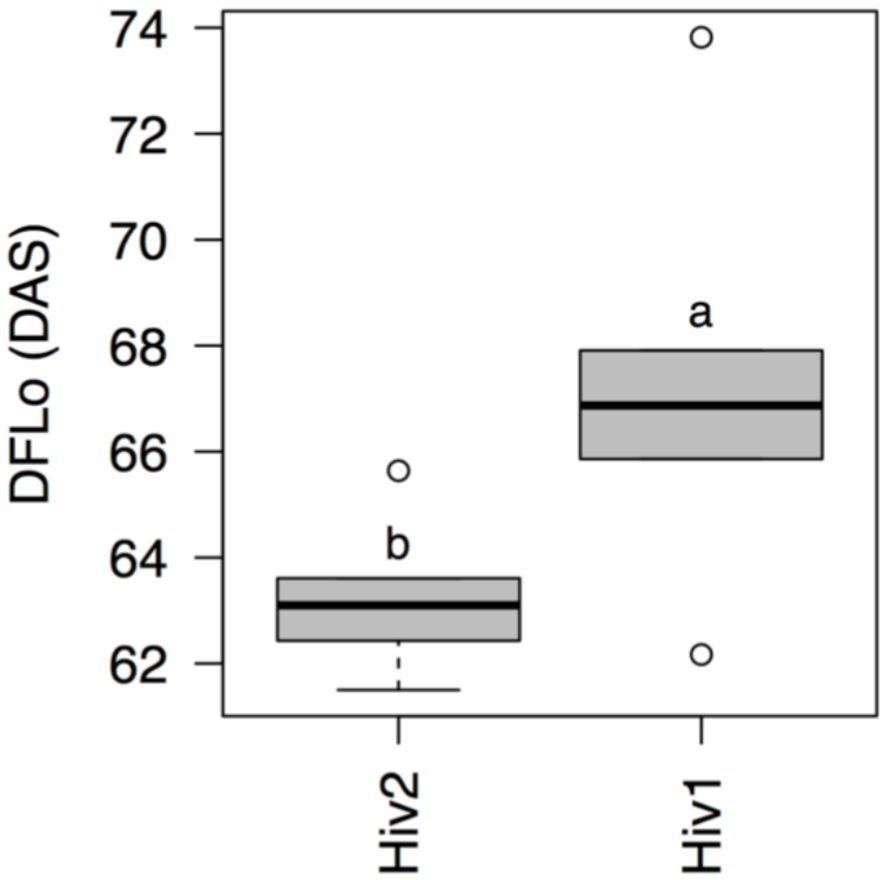



**Figure B- 3. Average performance values for days to flowering (DFLo) of check varieties within each experiment under rainfed conditions.** Different letters (e.g., a and b) indicate a significant difference between early (Hiv1) and late (Hiv2) planting date experiments based on the Tukey's Honest Significant Difference test. DAS means the number of days after sowing.

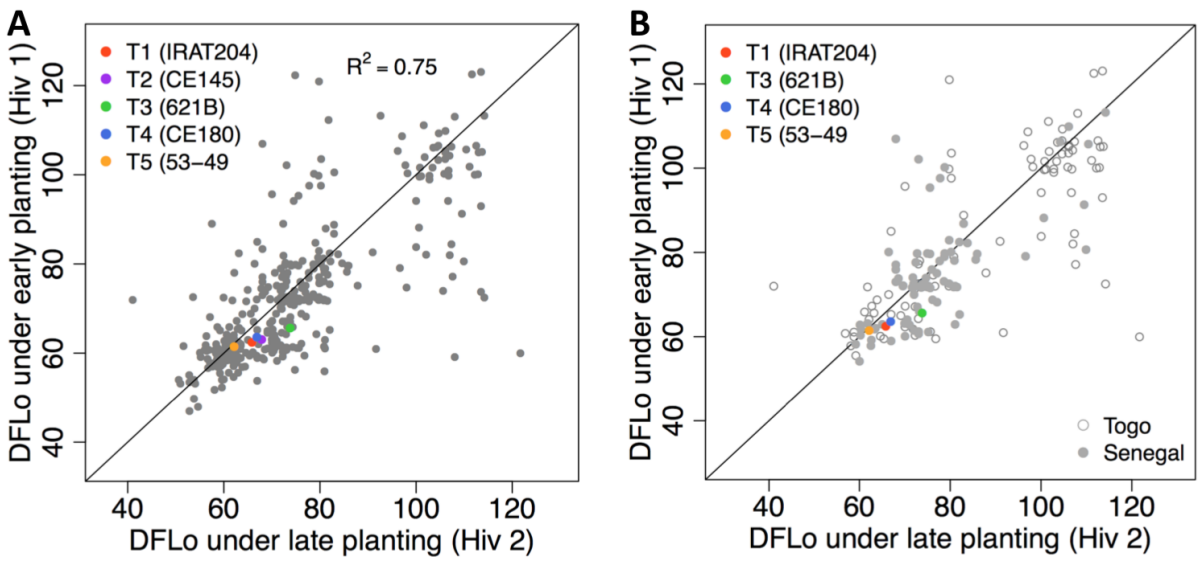



**Figure B- 4. Flowering time differences of accessions between early (Hiv1) and late (Hiv2) planting date experiments under rainfed conditions.** (A) Correlation for days to flowering (DFLo) between Hiv1 and Hiv2 within the whole WASAP. (B) Correlations for DFLo between Hiv1 and Hiv2 within Togo accessions (open circles) and within Senegal accessions (closed circles). The color-coded dots indicate the check varieties.

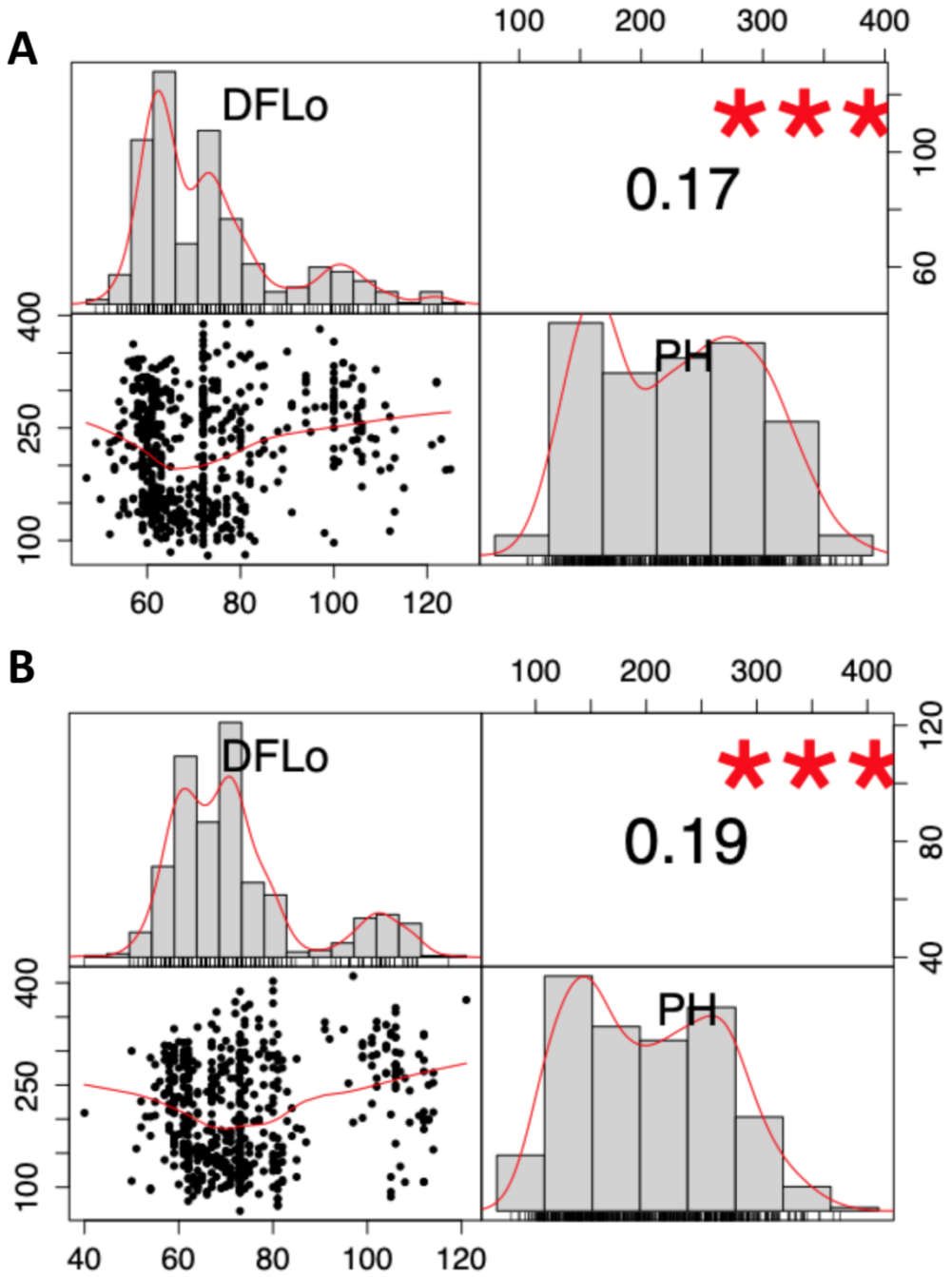



**Figure B- 5. Phenotypic correlations for DFLo, days to flowering and PH, plant height within (A) early (Hiv1) and (B) late (Hiv2) planting date experiments.**

**Table B- 1. Pairwise weighted $F_{ST}$ genetic differentiation among botanical types and countries of origin.**

| (A) **$F_{ST}$ among botanical types** | | |
|---|---|---|
| Botanical type | | $F_{ST}$ |
| Bicolor | Durra | 0.02 |
| Bicolor | Caudatum | 0.05 |
| Bicolor | Guinea | 0.11 |
| Bicolor | Gm | 0.07 |
| Bicolor | DC | 0.18 |
| Durra | Caudatum | 0.11 |
| Durra | Guinea | 0.14 |
| Durra | Gm | 0.11 |
| Durra | DC | 0.20 |
| Caudatum | Guinea | 0.15 |
| Caudatum | Gm | 0.14 |
| Caudatum | DC | 0.22 |
| Guinea | Gm | 0.04 |
| Guinea | DC | 0.18 |
| Gm | DC | 0.17 |
| *Average* | | 0.16 |
| (B) **$F_{ST}$ among countries of origin** | | |
| Country of origin | | $F_{ST}$ |
| Mali | Niger | 0.06 |
| Mali | Senegal | 0.01 |
| Mali | Togo | 0.12 |
| Niger | Senegal | 0.08 |
| Niger | Togo | 0.14 |
| Senegal | Togo | 0.12 |
| *Average* | | 0.09 |

DC, durra-caudatum types; Gm, guinea margaritiferum.

**Table B- 2. Pairwise weighted $F_{ST}$ among ADMIXTURE ancestral populations.**

| Populations | G-I | G-II | G-III | G-IV | G-V | G-VI | G-VII | Total $F_{ST}$ |
|---|---|---|---|---|---|---|---|---|
| **G-II** | 0.36 | | | | | | | |
| **G-III** | 0.35 | 0.43 | | | | | | |
| **G-IV** | 0.49 | 0.50 | 0.54 | | | | | |
| **G-V** | 0.31 | 0.25 | 0.43 | 0.61 | | | | |
| **G-VI** | 0.34 | 0.44 | 0.32 | 0.54 | 0.43 | | | |
| **G-VII** | 0.43 | 0.34 | 0.51 | 0.57 | 0.34 | 0.51 | | |
| **G-VIII** | 0.39 | 0.36 | 0.44 | 0.54 | 0.26 | 0.42 | 0.41 | 0.39 |

Accessions with >0.6 ancestry fractions for the given genetic group were included in $F_{ST}$ analysis.

**Table B- 3. Quantitative trait loci near *Ma6* and *SbCN8* candidate genes associated with days to flowering BLUPs using the GLM.**

| QTL[a] | *P*-value | MAF | Position (kb) | Locus name |
|---|---|---|---|---|
| S6_651847 | $< 10^{-23}$ | 0.25 | 45 | *Ma6* |
| S6_697299 | $< 10^{-18}$ | 0.24 | 160 bp | *Ma6* |
| S6_699842 | $< 10^{-18}$ | 0.22 | within | *Ma6* |
| S6_699843 | $< 10^{-18}$ | 0.23 | within | *Ma6* |
| S9_54917833 | $< 10^{-20}$ | 0.25 | 43 | *SbCN8* |
| S9_54968379 | $< 10^{-19}$ | 0.26 | 4 | *SbCN8* |

[a] Digit before and after underscore indicates chromosome number and SNP position on the genome, respectively; BLUP, best linear unbiased prediction; GLM, general linear model; QTL, quantitative trait locus; MAF, minor allele frequency.

**Table B- 4. Quantitative trait loci near *qHT7.1* and *Dw3* candidate genes associated with plant height BLUPs using the GLM.**

| QTL[a] | *P*-value | MAF | Position to locus (kb) | Locus name |
|---|---|---|---|---|
| S7_56232413 | $< 10^{-10}$ | 0.19 | 230 | *qHT7.1* |
| S7_56432423 | $< 10^{-10}$ | 0.20 | 30 | *qHT7.1* |
| S7_59955806 | $< 10^{-10}$ | 0.14 | 125 | *Dw3* |
| S7_59402662 | $< 10^{-9}$ | 0.15 | 419 | *Dw3* |

[a] Digit before and after underscore indicates chromosome number and SNP position on the genome, respectively; GLM, general linear model; QTL, quantitative trait locus; MAF, minor allele frequency.

**Table B- 5. Quantitative-trait loci associated with plant height BLUP across early and late planting date experiments using the MLMM.**

| QTL[a] | MLMM *p*-value | MAF | Effect size | PVE[b] (%) | MLM *p*-value | Position to locus (kb) | Locus name |
|---|---|---|---|---|---|---|---|
| S5_61867719 | $< 10^{-10}$ | 0.09 | 44 | 24.7 | $< 10^{-6}$ | | |
| S5_30001948 | $< 10^{-9}$ | 0.04 | 41 | 13.3 | 0.003 | | |
| S9_38942669 | $< 10^{-8}$ | 0.24 | 62 | 20.9 | $< 10^{-5}$ | | |
| S7_59400476 | $< 10^{-5}$ | 0.15 | 73 | 40.9 | $< 10^{-13}$ | 421 | *Dw3* |

[a] Digit before and after underscore indicates chromosome number and SNP position on the genome, respectively; [b] ADMIXTURE ancestry fractions at K = 8 were used as fixed effect covariates; BLUP, best linear unbiased prediction; MLMM, multi-locus mixed-linear model; QTL, quantitative-trait loci; MAF, minor allele frequency; PVE, proportion of variance explained; MLM, mixed-linear model.

**Table B- 6. List of GWAS QTLs using MLMM, excluding those at *Ma6*, *SbCN8*, and *Dw3* overlapping with published QTLs from other studies based on the sorghum QTL Atlas.**

| GWAS QTL[a] | QTL ID | LG:start–end | Original Reference |
|---|---|---|---|
| Days to flowering QTLs | | | |
| S4_57407080 | *QDTFL4.21* | 4:56.1–61.6 | (Felderhoff *et al.*, 2012) |
| S8_2206437 | *QDTFL8.5* | 8:1.9–2.7 | (Wang *et al.*, 2014) |
| | *QDTFL8.8* | 8:2.3–2.7 | (Mace *et al.*, 2013a) |
| S9_55345348 | *QDTFL9.16* | 9:50.4–57.9 | (Lin *et al.*, 1995) |
| | *QDTFL9.15* | 9:50.4–57.9 | (Feltus *et al.*, 2006) |
| | *QDTFL9.33* | 9:56.1–59.1 | (Zhang *et al.*, 2015) |
| S10_9523248 | *QDTFL10.10* | 10:7.6–9.7 | (Wang *et al.*, 2014) |
| S10_51083132 | *QDTFL10.26* | 10:42.1–51.9 | (Sangma, 2013) |
| Plant height QTLs | | | |
| S5_61867719 | *QHGHT5.11* | 5:62.3–62.9 | (Bouchet *et al.*, 2017) |

[a] Digit before and after underscore indicates chromosome number and SNP position on the genome, respectively; QTL, quantitative trait locus; MLMM, multi-locus mixed-linear model; LG, linkage group; SNP, single nucleotide polymorphism.

# Chapter 4 - Genome-Wide Association Studies of Drought Tolerance in West African Sorghum


## Abstract

Sorghum (*Sorghum bicolor*), a staple food crop in Africa, is affected by early and end season droughts in semi-arid regions. Dissecting the genetic architecture of yield stability under various drought scenarios can facilitate breeding to develop climate resilience varieties. Genome-wide association studies (GWAS) in locally adapted varieties can identify positive pleiotropic loci for yield stability across drought scenarios. This study aimed to identify environment-specific and positive pleiotropic drought-yield loci across 756 sorghum accessions of West Africa. The fraction of transpirable soil water in stress environments decreased to ~0.3 of that of field capacity, indicating that plants went through severe water limitation at different stages. Phenotypic variation was considerably influenced by genotype and genotype by environment interaction effects ($p < 0.01$). Broad sense heritability for yield components was moderate to high across environments. Significant correlations were observed for yield components in specific drought scenarios and across environments. Several genotypes performed better than the drought reference checks, B35 and Tx7000 in both pre-and post-flowering drought scenarios. Many lead associations (134) were commonly identified by general-linear model and mixed-linear model in GWAS analysis for reduction of yield components and stress tolerance index for grain weight in independent stressed environments. Some associations had high allelic effects and explained 11 to 27% of phenotypic variance. Twenty-nine lead associations colocalized with *Stg1–4* loci, which overlapped with signatures of positive selection to dry environments. Overall, this study contributes to understanding the genetic architecture of drought tolerance. Drought tolerance loci can facilitate improving locally preferred varieties while maintaining high productivity via in marker-assisted selection.


## Introduction

Unpredictable rainfall and drought scenarios occurring during growing seasons considerably decrease crop productivity semiarid regions. Improving crop adaptation to water limitation is critical for establishing food security in developing countries characterized by the vulnerability of smallholder farmers to climate changes (Mundia *et al.*, 2019). Drought tolerance is a complex phenomenon that affects several aspects of plant development and phenology

relative to yield potential (Blum, 2005). An understanding of the genetic architecture of grain yield and its components across various drought scenarios can facilitate crop breeding to increase production. Advances in genomic sequencing technologies and quantitative genetic mapping facilitate the dissection of genomic regions controlling agronomic traits to accelerate breeding through genomic-assisted breeding (Poland, 2015; Fu *et al.*, 2017). However, collecting good phenotypic data under well managed water stress environments and integrating phenotypes to genotypes remain a major constraint in many small breeding programs. The genetic dissection of yield components under various drought scenarios would provide favorable alleles for drought tolerance with high productivity.

Sorghum (*Sorghum bicolor*) is a staple cereal food crop in many developing countries but is mainly affected by water scarcity in drought-prone environments. Sorghum is a resilient model crop relatively tolerant to water deficit relative to many cereal crops (Mullet *et al.*, 2014). Studies have demonstrated that genetic variation controls drought tolerance in sorghum. Major effect dominant stay-green loci (*Stg1–4*) and lately *Stg5* underlie post-flowering drought tolerance in bi-parental populations and near-isogenic lines (Tuinstra *et al.*, 1997; Xu *et al.*, 2000; Harris *et al.*, 2007; Borrell *et al.*, 2014b; Hayes *et al.*, 2016). Drought tolerance loci in sorghum influence several aspects of sorghum development, including canopy architecture, water supply, phenology and grain yield (Borrell *et al.*, 2014b). These loci explained between 8 to 30% of the stay-green trait variation (Xu *et al.*, 2000; Harris *et al.*, 2007). However the existence of natural variants at these loci in West African locally-adapted varieties is poorly known. In addition, none of alleles at these loci have been identified in diverse genetic backgrounds in Sub-Saharan Africa. Understanding the genetic basis of drought tolerance in locally-adapted sorghum will contribute to enable breeding to rapidly develop drought tolerant varieties. The identification of favorable alleles for drought tolerance while maintaining high productivity is required in breeding. Stress tolerance index (STI) allows to assess genotypes with high production capacity under both stress and non-stress environments (Thiry *et al.*, 2016). STI has been successfully used to identify genetic associations with drought tolerance in association mapping studies (Li *et al.*, 2018; Yuan *et al.*, 2019).

Local varieties have been under natural and farmers selection for adaptation to various environmental conditions and farming systems. Although there might exist physiological tradeoffs between pre- and post-flowering drought tolerance, selection may have fixed favorable

large-effect loci early on (Orr, 1998). Local varieties are adapted to various environmental conditions since their domestication 8000 y ago followed by diversification (Harlan & De Wet, 1972; Wendorf *et al.*, 1992). Consequently, positive pleiotropic QTL for combined pre- and post-flowering drought tolerance might exist in locally adapted varieties. A positive pleiotropic QTL is defined here as a QTL that is associated with drought response in both pre- and post-flowering drought scenarios or multiple drought stress environments. Positive pleiotropic loci for tolerance to pre- and post-flowering drought scenarios are poorly established in breeding. West African sorghum is extremely diverse and there have been few cycles of selection in breeding programs (Mauboussin *et al.*, 1977; Leiser *et al.*, 2014). The West African sorghum association panel (WASAP) including landraces and breeding lines that consist of working collection of breeding programs was assembled and genotyped using genotyping-by-sequencing technology (chapter 3). The germplasm contains natural genetic variants favorable for adaptation to a wide range of environments. However, the genetic architecture underlying grain yield and its components under various drought scenarios remains largely unknown in the germplasm. We hypothesized that positively pleiotropic QTLs confer combined pre- and post-flowering drought tolerance in the West African sorghum.

Genome-wide association studies (GWAS) contribute to the identification of a large number of natural variants of known genes with high resolution by taking advantage of historical recombinations within diversity panels (Yu & Buckler, 2006; McCouch *et al.*, 2016; Yano *et al.*, 2016; Zhao *et al.*, 2019). Grass species such as sorghum are suitable to identify natural variants underlying complex agronomic traits partly due to its small genome size and moderate LD (Paterson *et al.*, 2009; Mace *et al.*, 2013; McCormick *et al.*, 2018). The sorghum QTL Atlas represents a useful resource to compare GWAS associations with drought tolerance QTLs identified from different studies based on linkage mappings (Mace *et al.*, 2019). Disentangling positive pleiotropic effects of drought-yield QTLs through GWAS can contribute to detect and characterize the natural allelic variation existing within locally-adapted populations. In this study, we performed GWAS on 756 sorghum accessions of the WASAP under ten different environments using previous GBS SNP dataset. We (*i*) characterize the genetic variation of yield components under various water stress environments; (*ii*) identify genetic variants at known and novel drought tolerance loci while maintaining high productivity under pre- and post-flowering water stress environments; (*iii*) investigate the pleiotropic effect of drought-yield QTLs

associated with STI and reduction of yield components under various drought scenarios; and (*iv*) determine signatures of selection overlapping identified drought-yield QTLs. The present study provides knowledge of the genetic architecture of yield components under various drought scenarios.

## Materials and Methods

### Plant materials and field experiments

The West African Sorghum Association Panel (WASAP) consists of 756 accessions from the four West African countries of Senegal (with 118 accessions genotyped), Mali (123), Togo (156), and Niger (359). The panel includes predominantly landraces grown by smallholder farmers across various environmental conditions and also breeding lines and improved varieties with useful agronomic traits. Five local breeding lines were used as checks, T1 (IRAT 204 or CE151-262), T2 (CE145-266), T3 (621B), T4 (CE180-33) and T5 (53-49). Two international drought tolerance lines, Tx7000 for pre- and BTx642 (formerly known as B35) for post-flowering drought tolerance were used as reference checks. For selection scans, we included 550 worldwide sorghum accessions including wild relative sorghum accessions with available sequencing data (Morris *et al.*, 2013).

Accessions were planted under field conditions for four years (2014, 2015, 2016, and 2017). Field experiments were performed at the Bambey Research Station, CNRA–Centre National de Recherche Agronomique (14.42˚N, 16.28˚W) in Senegal. The experiment site is located in the Soudano-Sahelian zone characterized by a short growing season, starting from July to October and annual precipitation < 600 mm. The monthly precipitation reaches its peak in August (Fig. 1). In total, ten experiments were performed in an incomplete randomized block design across the four years. Two adjacent experiments were carried out under rainfed conditions (RF) in 2014, RF1 and RF2 with one month planting date interval (August to December 2014). Two adjacent experiments, well water (WW), pre-flowering water stress (WS1) were planted during the dry hot-off season in 2015 (March to August). Three adjacent experiments, WW, WS1, and post-flowering water stress (WS2) were planted during the cool-off season in 2015-2016 and 2016-2017 (October 2015 to March 2016). Each experiment in a year is considered as an environment.

During the growing season of 2014, the cumulative rainfall recorded was 394.9 mm. The average daily temperature varied between 22.4 and 35 °C and average relative humidity between

42.14 and 89.27%. In WW, irrigation was applied twice a week (30 mm each time) until physiological maturity. In WS1, water stress was applied 30 DAP to mimic a one month early season drought and irrigation was restarted 60 DAP until physiological maturity. In the WS2, water stress was applied when 75% of plants in a maturity group flowered and was maintained until physiological maturity. Three maturity groups were defined based on accessions phenology characterized during 2014 experiments for water deficit application in WS2. In each environment, phenological, physiological, and yield component traits were measured.

**Agronomic measurements**

Days to 50% flowering (DFLo) of plants in a plot (one row), above ground biomass (DBM), plant height (PH), and yield related traits including grain weight per panicle (GrW), panicle weight (PW), grain number per plant (GrN), and thousand grain weight (TGrW) were measured and used for association mapping studies. For each trait except for DFLo and TGrW, three plants from each plot were used for measurements. The drought stress tolerance index (STI) for grain weight was calculated from the GrW under WW and WS1 or WS2 as follows:

$$STI = \frac{(Y_{ww})(Y_{ws})}{Y^2_{m.ww}}$$

Where *Yww* and *Yws* is the grain weight of a given genotype in control environments, respectively and *Ym.ww* is the mean value of GrW in the control environment. For the STI, the higher the value, the more tolerant the genotype to the stress. The drought reduction of each yield component relative to the control environment was calculated as follow:

$$Ri(\%) = \frac{Y_{ww} - Y_{ws}}{Y_{ww}} \times 100$$

Where *Ri* is the drought response of a genotype for a trait *i*, *Yww* and *Yws* are the performance of the genotype in control environment and water stressed environment, respectively.

**Statistical analysis and phenotypic evaluation**

Statistical analysis was performed using the R program (R Core Team, 2016). The variance components were estimated by fitting the mixed linear model with random effects for all genotypes (G), water regimes (WR), years (Y), and G x Y interaction effects using the *lme4* package (Bates et 2010). Broad sense heritability ($H^2$) was calculated based on variance components derived from the mixed effect model. Heritability was estimated for each trait across environments based on the genotypic variance and the total phenotypic variance. Phenotypic correlations among traits were calculated using Pearson correlation of the PerformanceAnalytics

package (Peterson *et al.*, 2014). Tukey's Honest Significant Difference test of the Agricolae package (Mendiburu, 2009) was used to test the difference of genotype performance between environments or botanical types. The BLUP values of the phenotypes were calculated by combining data for a given water regime across years. The phenotypic BLUPs were used for the genome-wide association analysis.

## Genome-wide association studies

To identify drought-yield QTLs, GWAS was performed using the general linear model (GLM) with principal component eigenvalues and the mixed linear model (MLM) in *GAPIT* package (Lipka *et al.*, 2012). These two GWAS models were used as complementary because the GLM may identify false positive associations while MLM may lead to false negative associations when controlling for false positive associations. To reduce false positive associations due to low allele frequency polymorphisms, the SNP dataset was filtered again for MAF >0.02 (because rare variants can contribute to phenotypic variation). The five first principal components and kinship matrix were used to account for population structure and genetic relatedness effects, respectively for the MLM. The significance level of GWAS associations were defined based on Bonferroni-corrected *p*-value 0.05 for the GLM with PCA (termed as GLM along the text) or at least top five SNPs above $p < 10^{-5}$ cutoff for the MLM. The lead peak SNP within a 150 kb surrounding genomic region was chosen to represent the associated region. To verify LD around GWAS peaks, LD heatmaps of QTLs region or 300 kb region surrounding lead SNPs in QTLs were constructed using the R package *LD heatmap 0.99-4* (Shin *et al.*, 2006). The proportion of phenotypic variance was estimated using linear models with fractions of ancestry inferred by ADMIXTURE (Alexander *et al.*, 2009) used as fixed covariates to account for background effects. BLUP values of phenotypes across water stress environments were used for the estimation of the proportion of phenotype variance explained by GWAS lead SNPs.

## Genome-wide selection scans

Genome-wide selection scans were performed to assess the impact of selection on GWAS associations based on 100-kb sliding windows using the vcftools program (Danecek *et al.*, 2011). Decreased in genome-wide nucleotide diversity was determined in Niger, Senegal, and Mali accessions relative to Togo accessions. Colocalization between significant associations and

domestication/improvement selective sweep regions were identified. The direction of selection of potential polymorphism at the *Stg3a* locus was determined using Tajima's D test.

## Results

### Phenotypic variation in the WASAP

About 600 accessions of the West African sorghum association panel (WASAP) were evaluated for phenological, physiological, and yield component traits under ten environments during four years at Bambey in Senegal. Average monthly precipitation and temperature at the field location reflect the Sahelian-Soudanian climate variation with the maximum precipitation obtained in August (Fig. 4-1A, B). Day length varies about one hour and half across the year (Fig. 4-1B). To test that water deficit was precisely controlled for drought phenotyping in managed water stress conditions, we estimated the fraction of transpirable soil water (FTSW) in the WS1, WS2, WW different environments. FTSW was estimated to be 0.6% in control and stressed environments before water stress application (Fig. 4-1C). FTSW then decreased considerably to ~0.2 and 0.3 in WS1 and WS2 environments, respectively. In both WS1 and WS2, plants went through severe water stress at different stages of their development depending on the environment.

To confirm the effect of water deficit at different stages of plant development, we determined the grain yield and days to flowering of genotypes in each water regime and environment. As expected, a strong G x E interaction was observed between the two drought tolerance reference lines, B35 and Tx7000 in WS1 and WS2 (Fig. 4-1D). Their average grain weight was lower than that of the local drought reference check, CE145-266 in control environments. Average grain weight was generally reduced in stressed environments relative to control environments (Fig. 4-1E). However, average grain weight in WS1 of 2015 was not significantly different from control environments. DFLo of genotypes was significantly delayed in 2015 dry hot-off season environments, whereas it was reduced in cool-off seasons of 2016 and 2017 relative to rainfed conditions (Fig. 4-1F). Average grain weight was not significantly different between rainfed conditions (RF) and WS2 environments. Average grain weight was significantly different between RF and WS1 environments (Fig. 4-2A). As expected, the reduction of GrN was significantly higher in WS1 (Fig. 4-2B).

To determine that the phenotypic variation in the WASAP, yield components were analyzed under WS1, WS2, RF, and WW water regimes and across all environments. A

considerable phenotypic variation was observed in the panel ($p < 0.01$) with coefficients of variation varying from 6% for DFLo in RF to 62% for GrW in WS1 environments across years (Table C-1). DFLo was delayed in WS1, whereas it was not different in WS2 relative to the WW controls. Significant genotypic variance and genotype by environment interaction effects were observed across environments for each trait (Table 4-1). To assess the part of the phenotypic variation that is due to genetic differences in the panel, the broad-sense heritability ($H^2$) of yield components was estimated from the estimated genetic and residual variances. As expected, $H^2$ estimates varied from moderate to high with values ranging from 0.53 for GrN to 0.95 for DFLo and PH (Table 4-1).

In sorghum breeding, the caudatum type are thought to have higher yield advantage over guinea and durra types. Unexpectedly, the average grain weight was not significantly different between caudatum accessions and durra and guinea accessions within each water regime (Fig. 4-2C). Durra-caudatum intermediates had higher average grain weight than caudatum accessions. Given that durra type is more adapted to drought conditions, we predicted that the reduction of average GrW and GrN values in water stressed environments would be significantly less in durra accessions than other types. Reduction of GrW values was not significantly different between durra accessions and caudatum and guinea accessions within each water regime (Fig. 4-2D). The durra-caudatum intermediates had higher average grain weight than durra accessions. The reduction of GrW and GrN was significantly higher in WS1 than in WS2 (Fig. 4-2D, F).

**Phenotypic correlations in specific and across water stress environments**

To determine whether there would be strong positive correlation among yield components WS1 and WS2 or across water stressed environments, Pearson's correlation coefficients were estimated. Significant correlations were observed among yield components under the same WS1 or WS2 (Fig. 4-4A). However, there was no or weak correlation of phenotypes between WS1 and WS2, except for GrW, DBM, and STI for grain weight. High positive correlation was observed between the BLUP values of GrW, PW, DBM, and GrN, while thousand grain was negatively correlated with grain number (Fig. 4-3). To verify if some genotypes performed better than the drought tolerance reference check lines in both WS1 and WS2 of 2016 and 2017, we determined the 1:1 ratio for grain weight and STI for grain weight. As expected, some genotypes performed better than the reference check lines in both water stress environments over the two years (Fig. 4-4). Overall, there was considerable phenotypic variation

in the panel, with a significant genetic component contributing to this variation across different environments.

## Genome-wide association studies of flowering time

To identify loci underlying drought tolerance, we carried out GWAS analyses using 130,709 SNP markers. Firstly, we determined that the phenotypic data are of sufficient quality for effective GWAS analyses. We used DFLo under control environments of 2015, 2016 and 2017 to map known flowering time candidate genes using the GLM with principal component analysis (PCA) to account for population structure effect. Flowering time is a highly heritable trait and its genetic control has been characterized in sorghum, maize, and rice. Any significant peak above the Bonferroni-corrected *p* value of 0.05 was identified by GLM for DFLo of the 2015 data (Fig. 4-5). Significant associations were identified for DFLo of the 2016 and 2017 data (Fig. 4-5). The SNPs, S6_55280640 and S3_62811196 were significantly associated with DFLo in both years. These two SNPs co-localized with *Zfl1* (9 kb away) and *SbCN12* (61 kb away) flowering time candidate genes, respectively. In both years, S6_55280640 was the lead SNP (*p* value < $10^{-10}$ in 2016; *p* value < $10^{-10}$ in 2017) of the associated region on chromosome 6. A third SNP, S2_67812515 was significantly associated with DFLo in 2017 data and colocalized with *Ma2* candidate gene (70 kb away). When the MLM with PCA and kinship were used to control for background effects, significant associations were not identified at the Bonferroni correction threshold. The association between S6_55280640 and DFLo in 2017 (*p* value < $10^{-5}$) was below the threshold.

## Associations for drought tolerance in independent water stressed environments

To determine the genetic natural variants associated with drought tolerance in the West African sorghum, we identified GWAS associations for reduction of PW (RPW), DBM (RDBM), GrN (RGrN), PH (RPH), TGrW (RTGrW), and STI for grain weight in specific water stressed environment and across environments. In WS1, reduction of DBM, GrN, and PH was severe in the pre-flowering drought stress environment. Whereas, the reduction of DBM and TGrW was severe in the post-flowering drought stress environment. These traits were used to identify associated variants at known drought tolerant loci in WS1 and WS2 environments, respectively, in addition to RPW and STI for grain weight. In total, 222 and 214 associations were identified by the GLM and MLM, respectively for drought response variables and STI for

grain weight in independent water stress environments. Among the associations, 134 were commonly identified by both GWAS models.

To identify positive pleiotropic effect QTLs between pre-and post-flowering water stressed environments, common associations in different water stressed environments were determined based on the 134 associations. A pleiotropy QTL was defined as one SNP or locus being mapped in both pre- and post-flowering drought scenarios, different stress environments, or associated with several drought response variables. As expected, many pleiotropic SNPs for drought response variables were observed across water stress environments (Table 4-2). For example, the SNP S8_58355080 on chromosome 8 was associated with STI in WS1 of 2016 and 2017 and in WS2 of 2016 and 2017 in both GLM and MLM models. The SNP S4_67777846 on chromosome 4 was associated with STI under WS1 of 2016 and 2017 and WS2 of 2017 in both GLM and MLM models. The S3_13763609 on chromosome 3 and S1_74186408 on chromosome 1 were associated with RPW in WS1 and WS2 of 2017 in both GLM and MLM. The identified pleiotropic SNPs showed significant allelic effect and significantly ($p < 10^{-8}$, the least highest value) explained 11 to 25% phenotypic variation for productivity under drought, with an average of 17% (Table 4-2).

**Drought response associations colocalizing with drought tolerant loci**

To verify the existence of natural variation at known drought tolerant loci involved in yield stability in the West African sorghum germplasm, we determined the colocalization between drought response-associations and known stay-green loci *(Stg1–4)*. Twenty-nine lead SNPs significantly associated with STI for grain weight and drought response variables colocalized with the known *Stg1*, *Stg2*, *Stg3a*, *Stg3b*, and *Stg4* loci (Table 4-3). The lead SNPs at each *Stg* locus explained 16% ($p < 10^{-10}$, *Stg1*), 20% ($p < 10^{-13}$, *Stg2*), 19% ($p < 10^{-13}$, *Stg3a*), 27% ($p < 10^{-16}$, *Stg3b*) and 21% ($p < 10^{-15}$, *Stg4*) of the phenotypic variance across WS1 and WS2 based on STI BLUP values. Each *Stg* locus showed positive pleiotropy by covering several associations for different drought scenarios and drought-yield response variables.

The *Stg2* locus covered pleiotropic associations for STI in WS1 of 2015 and 2017, WS2 of 2017, RGrN in WS1 of 2017 and RDBM in WS2 of 2016. There was a strong LD between several of the lead SNPs at the locus (Fig. 4-6A). The *Stg1* locus covered associations for RPW in WS1 and WS2 of 2017 and associations for STI in WS1 of 2017. There was a strong LD among lead SNPs within the locus (Fig. 4-6A). The *Stg3a* and *Stg3b* (which are next to each

other) region covered pleiotropic associations for STI in WS1 of 2015 and 2016, STI in WS2 of 2015 and 2017, RPW in WS1 of 2015 and 2017, and RDBM in WS2 of 2016. There was a strong LD among the lead SNPs within *Stg3b*. There was no LD among lead SNPs within *Stg3a* (Fig. 4-6B). The *Stg4* locus covered associations for RPW in WS1 of 2017 and for STI in WS1 of 2015 and in WS2 of 2017. In this locus, there was a moderately high LD between lead SNPs, S5_15916423 associated with RPW in WS1 and S5_52255304 associated with STI in WS2 (Fig. 4-6C). In the *Stg2* locus, the SNP S3_56094063 was the top significant association ($p < 10^{-19}$ in GLM and $p < 10^{-13}$ in MLM) for STI in WS2 and WS1. In the *Stg3b* locus, the SNP S2_62095163 was the top association ($p$, GLM $< 10^{-18}$ and MLM $< 10^{-13}$) with high effect for STI in WS2. This SNP was in strong LD with other lead SNPs in *Stg3b* (Fig. 4-6B) but not in LD with lead SNPs in *Stg3a*. Two lead SNPs, S2_60973403 and S2_59237127 only were in moderate ($r^2 < 0.3$) and low ($r^2 < 0.1$) LD with a non-synonymous SNP, S2_61595689 in *Stg3a* (Fig. 4-6D).

**Genome-wide selection signatures around drought response QTLs.**

We investigated the occurrence of positive selection in drought response lead SNP associations at *Stg1–4* loci for drought tolerance in Sahelo-Soudanian sorghums (Niger, Mali, and Senegal sorghums) relative to Guinean sorghums (Togo sorghums) based on ratios for decreased pairwise nucleotide diversity (*pi*). Selection outliers (genomic regions in the 95 percentile were considered as selection outliers) were identified within *Stg3a*, *Stg3b*, and *Stg4* in Niger and Senegal accessions (Fig. 4-7A, B). In Mali accessions, selection outliers were identified within *Stg3b*, *Stg2*, and *Stg4* (Fig. 4-7C).

To investigate the effect of domestication on drought response lead SNP associations, we assessed decreased *pi* in durra-caudatum (D-C) and durra landraces predominant in drought-prone areas compared to guinea landraces predominant in high rainfall areas. Selective sweep outliers at the 95 percentile were found in the durra-caudatums and durra relative to wild relative sorghums (Fig. 8A, B) and in improved lines relative landraces (Fig. 8C). Twelve lead SNP associations overlapped with domestication selective sweep outliers (Table 4-4). In durra-caudatums, selective sweep outliers were localized within *Stg1* (64.9–67.3 Mb), *Stg3a* (56.2–58.2 Mb), *Stg3b* (64.8–65.0 Mb and 68.3–68.4 Mb), and *Stg4* (13.6–19.0 Mb) (Fig. 8A). In landraces, selective sweep outliers were localized within *Stg1* (65.5–67.6 Mb), *Stg3b* (66.8–66.9 Mb), and *Stg4* (14.4–21.4 Mb) (Fig. 8B). In improved lines, selective sweep outliers were

localized within *Stg1* (at 62.6–63.2 Mb), *Stg3b* (71.2–71.4 Mb), and *Stg4* (13.1–16.6, 19.0–19.6 and 21.3–21.5 Mb) (Fig. 8C).

The selective sweep at 64.9–67.3 Mb in *Stg1* locus harbored the lead SNPs, S3_65137990, S3_65430305, S3_66366589, S3_66738018. The common allele frequency (CAF) at each of these lead SNPs was fixed (0.5; 100%) in durra-caudatums (Table 4-4). The selective sweep at 56.2–58.2 Mb in *Stg3a* harbored the lead SNP, S2_56682379 where the CAF was nearly fixed (0.48) in durra-caudatums. The selective sweep at 64.8–65.0 Mb in *Stg3b* harbored the lead SNP, S2_65658140 where the CAF was nearly fixed (0.48) in durra-caudatums. The selective sweep at 13.6–19.0 Mb in *Stg4* harbored the lead SNPs, S5_15215761, S5_15916423, and S5_16480120 with CAF fixed or nearly fixed (0.5, 0.5, and 0.45, respectively) in durra-caudatums. Selective sweep outliers at the 99 percentile were found only within *Stg1* and *Stg4* in durra-caudatums, *Stg4* in landraces, and *Stg3b*, *Stg1* and *Stg4* in improved lines.

### Haplotype associations at *Stg1* and *Stg3b* quantitative trait loci

To verify evidence of specific haplotype associations at *Stg1* and *Stg3b* for drought tolerance in durra-caudatums and improved lines, we analyzed haplotype associations with STI for grain weight across water stress environments based on BLUP values. Significant associations were identified at Bonferroni correction 0.05 between haplotypes based on all SNPs within *Stg1* for STI across the whole WASAP panel (Fig. 9A; Table 4-5). The top associated haplotype was formed by the SNPs, S3_64606653, S3_64612170, and S3_64623404, covering a 16.7 kb region. This haplotype block contained four haplotype alleles, among which the most significant association, AGC was associated with increased STI (beta, 3.5%; $p < 10^{-13}$) and was observed in 63% of durra-caudatums, while absent in durras, guineas, and improved lines. The allele CAT was significantly associated with decreased STI (beta, -2.3%; $p < 10^{-9}$) and observed in 33% of durra-caudatum, 85% of durras, 95% of guineas, and 79% of improved lines.

The second most significantly associated haplotype formed by the SNPs, S3_66396518 and S3_66396621 contained three haplotype alleles. The allele TA, which was the most significant, was associated with increased STI (beta, 2.1%; $p < 10^{-10}$) and observed in 76% of durra-caudatums, 61% of durras, 15% of guineas, and 27% of improved lines. The allele CT, significantly associated with decreased STI (beta, -2%; $p < 10^{-9}$) was observed in 23% of durra-caudatums, 39% of durras, 83% of guineas, and 73% of improved lines. Any of the lead SNPs

overlapped with top significantly associated haplotypes at *Stg1*. Also, there was no overlapping between lead SNPs with associated haplotypes defined based on non-synonymous SNPs at *Stg1* (Fig. 9B).

Significant haplotype associations to STI were found at the *Stg3b*, with four haplotypes containing four S2_62973945, S2_63881780, S2_65658140, and S2_69575903 (Fig. 9C; Table 4-6). The top associated haplotype was formed by 7 SNP covering a 7.1 kb region (70,827,637–70,834,834 bp). Haplotype allele TGGATGA, significantly associated with increased STI (beta, 3.1%; $p < 10^{-12}$) was observed in 67% of durra-caudatums, 7% of durras, 0% of guineas, and 3% of improved varieties. Whereas, haplotype allele TCGACTT, significantly associated with decreased STI (beta, -2%; $p < 10^{-10}$) was observed in 21% of durra-caudatum, 46% of durras, 81% of guineas, and 83% of improved varieties. The second most significantly associated haplotype was formed by 5 SNPs covering a 22.4 kb region [67,597,045–67,619,513 bp]. The haplotype allele GTACT, significantly associated with increased STI (beta, 3.1%; $p < 10^{-12}$) was observed in 67% of durra-caudatums, 0% of durras, 2% of guineas, and 3% of improved varieties. Whereas, haplotype allele CCTTG, significantly associated with decreased STI (beta, -1.8%; $p < 10^{-8}$) was observed in 9% of durra-caudatum, 22% of durras, 67% of guineas, and 3% of improved varieties. Four significantly associated haplotypes formed based on non-synonymous SNPs within *Stg3b* overlapped with the four lead SNPs (Fig. 9D; Table 4-6).

## Discussion

In the semiarid regions, growing seasons are characterized by drought, particularly early and end season droughts resulting in considerable yield loss (Mundia *et al.*, 2019). In this study, we performed GWAS to disentangle the existence of drought tolerance loci in the West African sorghum germplasm grown in common garden managed water experiments. Many positively pleiotropic drought response QTLs were identified, several of which colocalized with *Stg1-4* loci. Some of them overlapped with positive selection outliers in accessions from the semiarid climate, indicating the existence of *Stg* alleles in the West Africa sorghum germplasm, useful to facilitate drought tolerance sorghum breeding.

### Genetic differences contribute to phenotypic variation in the germplasm

Drought stress effect occurred in water stressed environments relative to control environments. The application of water deficit reduced the fraction of transpirable soil water (FTSW) to that of field capacity. The values obtained for FTSW were similar to the reported

values (0.26 to 0.37) in woody species for transpirable water stocked in the plant tissues to start (Sinclair *et al.*, 2005). This observation indicates that water supply was precisely controlled in the different environments to mimik drought scenarios that usually occur during the growing season. The effect of drought on plant performance was confirmed by the performance of the pre-flowering (Tx7000) and post-flowering (B35) drought tolerant reference lines as shown by the strong cross G x E interaction. As predicted, water deficit application resulted in significant reduction of grain yield and grain number and caused a significant delay of flowering time in water stressed environments. However, the reduction of grain yield in the water stress environment of 2015 was not significantly different from control environment of 2016. This variation occurred because the experiment was planted during the dry hot-off season (March to August), which is characterized by long growing days. The maturity cycle of genotypes was delayed (Fig. 4-1F), which allowed higher grain filling.

The phenotypic variation observed in the WASAP was highly influenced by genetic differences. Our prediction that water deficit significantly reduces grain number under pre-flowering drought relative to control environments held as the average grain number was significantly reduced in WS1. This reduction of GrN in WS1 is in line with the model that grain number is quantitatively affected by pre-flowering water stress. Under pre-flowering drought, fewer seeds during grain filling could increase photosynthate allocation to the seeds, resulting in improved fitness (likelihood of reproductive success). We also predicted that the average grain yield performance of the accessions would not be significantly different between water stressed environments in the cool-off season and rainfed conditions. There was not a significant difference between rainfed conditions and WS2 environments for grain yield. However, a significant difference was observed between RF and WS1 environments. This result suggests that post-flowering drought is more frequent and may have a higher impact on grain yield reduction during the growing season.

The high correlation between STI for grain weight and Plant height and dry biomass under pre-flowering drought stress suggests common genetic architecture in biological pathways involve in drought tolerance. A delay of flowering time was observed for the post-flowering drought tolerant line, B35 in stress conditions compared to Tx7000 across environments. Flowering time was not correlated with STI for grain weight under pre-flowering drought stress but was significantly correlated with STI under post-flowering drought. This significant

correlation suggests that flowering is a post-flowering drought escape adaptation. We evaluated the yield advantages among botanical types under different drought scenarios for potential use in breeding in West Africa integrating different types. The durra-caudatum types showed yield advantages under drought while guinea types showed more stability across environments. Integrating durra-caudatum intermediates and guinea sorghums in pre-breeding programs would be valuable as they show good adaptability under harsh environments. Hybrids from guinea sorghum in the Mali hybrid breeding program have been shown to harbor beneficial characteristics that meet farmer preferences (Kante *et al.*, 2017).

**Putative GWAS QTLs for productivity and drought response**

Drought tolerant and flowering time loci contribute to the genetic variation in the sorghum germplasm. We first showed the effectiveness of the association between phenotypic and genotypic data in GWAS using days to flowering as control to mapped known flowering time loci. We hypothesized that an oligogenic genetic architecture controls flowering time in the West African sorghum germplasm. We did not map the well known photoperiodic flowering genes in sorghum (*Ma1–Ma6*) as we expected. In contrast, we mapped two flowering time candidate loci, *Zfl1* and *SbCN12* in two consecutive cool-off seasons, which is characterized by short days. Sorghum is a shor day plant, such that it can flower anytime during the cool-off season regardless of photoperiod sensitivity. *Zfl1* was at the top GWAS peak. This gene has been hypothesized to act at the extreme of the flowering-time pathway in maize (Romero Navarro *et al.*, 2017), suggesting that *Zfl1* expresses during floral transition under short days without requiring a downregulation of other photoperiodic flowering genes in the pathway. This gene has similar activities to the orthologous of rice, *RLF* in inducing early flowering time during vegetative-to-reproductive transition (Rao *et al.*, 2008). In addition, there was no GWAS peak found in our 2015 dry hot off-season (long days) flowering time data characterized by long delay of or no flowering (Fig. 4-1F) before the critical photoperiod. This long delay of flowering suggests that *Zfl1* may not be significantly expressed. The minor allele frequency at *Zfl1* QTL was high (0.47) in the WASAP, indicating that most genotypes will have *Zfl1* functional allele.

Positive pleiotropic QTLs/genes showing stability across water stress environments are of potential interest for use in breeding through genomics-assisted selection. We carried out GWAS on drought responses relative to control environments to identify independent and positive pleiotropic loci for different water regimes. Reduction of yield components under water stress

relative to control environments were used as drought response variables to identify drought tolerance associations. The stress tolerance index for grain weight was used to identify associations for drought tolerance with high production (Thiry *et al.*, 2016). As we predicted, several drought response associations, were identified repeatedly in different water stressed environments and in different years (Table 4-2). This contrasts the competing hypothesis that there is no positive pleiotropic SNP for pre-flowering and post-flowering droughts. Although some of the associations could be related to background effects, several putative major effect QTLs were consistent across several environments and significantly contributed to phenotypic variation. The average proportion of phenotypic variation explained by lead SNPs at each *Stg* locus is comparable to the ~10 to 30% phenotypic variation explained by known stay-green loci (Xu *et al.*, 2000; Harris *et al.*, 2007). However, the phenotypic variation explained by *Stg3b* QTL was the highest (27%) in the WASAP accessions, in contrast to the known stay-green sources where *Stg2* had the largest contribution to the phenotype variation among the *Stg1–4* loci (Xu *et al.*, 2000; Harris *et al.*, 2007). This result might be explained by interaction of *Stg* QTLs with other genetic factors in this population composed of diverse unrelated accessions compared to the near-isogenic lines (NILs) used in previous studies.

The findings support our favored hypothesis that positively pleiotropic QTLs contribute to combined drought tolerance to multiple environments in the West African sorghum. Putative QTLs were colocalized with most of the drought tolerant loci from the sorghum QTL Atlas (Mace *et al.*, 2019), demonstrating the power of GWAS to dissect the genetic basis of drought tolerance in the locally adapted sorghum germplasm. The identification of all the *Stg1–4* coupled with the significant PVE by loci supports our hypothesis that these loci exist also in the West African germplasm, beyond the known sources of stay-green in sorghum (Tuinstra *et al.*, 1997; Haussmann *et al.*, 2002; Harris *et al.*, 2007; Borrell *et al.*, 2014; Hayes *et al.*, 2016).
Natural variation in drought tolerant loci shows allelic differentiation between semi-arid and humid zones in West Africa. Signatures of positive selection overlapped with *Stg4* in Senegal, Niger, and Mali accessions and *Stg3a* in Senegal and Niger accessions. Stay-green may have been selected for drought adaptation in the Sahel versus the humid climate, consistent with high frequency of haplotypes within *Stg1* and *Stg3b* associated with increased drought tolerance in durra-caudatums. The *Stg2* QTL may have been selected in Togo accessions during domestication to confer drought adaptability in dry areas. The characterization of *Stg1–4* alleles

effects in diverse backgrounds and environments would shed more lights on the expression of stay-green in sorghum across diverse environments in West Africa. Regarding our hypothesis that novel drought tolerant loci contribute to yield stability in the West African sorghum, several novel associations for STI were identified in the germplasm. Some of the associations did not colocalize with known drought tolerance loci, suggesting that novel drought tolerant loci may contribute to drought adaptation of West African sorghum. The development of near-isogenic lines (NILs) and the validation of major effect QTLs in breeding populations will be crucial to characterize the interaction of these QTLs with environmental factors and different elite backgrounds in West Africa. The drought response QTLs were identified based on phenotypic responses at the plant level rather than responses per unit land area. Traits such as PH, DBM, PW, GrW, and GrN might be the most sensitive to the effects of variable plant density compared to DFLo. It would be interesting to investigate grain yield responses of the identified QTLs, in different backgrounds, to water availability in regard to plant density for both response/plant and response/$m^2$ (Dhungana *et al.*, 2007; Houshmandfar *et al.*, 2019).

The GWAS analysis was effective in detecting positive pleiotropic loci underlying multiple traits under various environments. GWAS analysis detected several genomic regions harboring lead SNPs associated with multiple drought response variables. The *Stg1–4* loci harbored several lead SNPs associated with multiple response variables and STI for grain weight in different environments. This observation is consistent with the role of *Stg* loci on grain yield stability, phenological and morphological phenotype variation in sorghum (Borrell *et al.*, 2014b,a). A major effect QTL led by the SNP S10_4711152 on chromosome ten was associated with RPW and RGrN in WS1 of the 2015 dry hot-off season environment with longer growing cycles. Phenotypic variation in this WS1 2015 environment was comparable to all control environments in this study, suggesting that this QTL may underlie similar genetic mechanisms for panicle weight and grain number under well water conditions rather than pre-flowering water stress conditions. Grain yield was significantly lower in the WS1 drought scenario. Most lead SNPs associated with RPW were detected under WS1 environments, suggesting a critical impact of early season droughts in yield reduction in the semiarid regions. Study in pearl millet in Sub-Saharan Africa also revealed a considerable impact of early season droughts on yield reduction (Debieu *et al.*, 2018), suggesting the need to investigate physiological and genetic mechanisms

underlying effects of water deficit on yield reduction of plants from vegetative to maturity stages.

## Conclusion

We demonstrated that the phenotypic variation in the WASAP is due to genetic differences among genotypes, with a considerable genotype-environment interaction. The present study has contributed to the understanding of the genetic basis of grain yield under water limitation. Large-effect positive pleiotropic QTLs were identified in multiple water stress environments. The identified loci significantly contributed to the phenotypic variation. This study demonstrates that the genetic architecture of stay-green is oligogenic and that *Stg1–4* alleles exist in the genetic background of West African sorghum. *Stg3a*, *Stg3b*, *Stg4* overlapped with signature of positive selection to semiarid regions. This result indicates the potential interest of *Stg1–4* alleles in marker-assisted selection to improve for drought tolerance, which is difficult to achieve using phenotypic selection. The natural variants that contributed the most to the phenotypic variance can be converted into high-throughput breeder-friendly markers to follow the introgression of *Stg* alleles into elites West African backgrounds. Near-isogenic lines at these loci will contribute to understanding the genetic mechanisms underlying yield stability, the effect of *Stg1–4* alleles in elite backgrounds, and their interaction across multiple environments in West Africa.

# References


Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* 19: 1655–1664.

Borrell AK, Mullet JE, George-Jaeggli B, van Oosterom EJ, Hammer GL, Klein PE, Jordan DR. 2014a. Drought adaptation of stay-green sorghum is associated with canopy development, leaf anatomy, root growth, and water uptake. *Journal of Experimental Botany* 65: 6251–6263.

Borrell AK, van Oosterom EJ, Mullet JE, George-Jaeggli B, Jordan DR, Klein PE, Hammer GL. 2014b. Stay-green alleles individually enhance grain yield in sorghum under drought by modifying canopy development and water uptake patterns. *New Phytologist* 203: 817–830.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, *et al.*2011. The variant call format and VCFtools. *Bioinformatics* 27: 2156–2158.

Debieu M, Sine B, Passot S, Grondin A, Akata E, Gangashetty P, Vadez V, Gantet P, Foncéka D, Cournac L, *et al.*2018. Response to early drought stress and identification of QTLs controlling biomass production under drought in pearl millet. *PLOS ONE* 13: e0201635.

Fu Y-B, Yang M-H, Zeng F, Biligetu B. 2017. Searching for an Accurate Marker-Based Prediction of an Individual Quantitative Trait in Molecular Plant Breeding. *Frontiers in Plant Science* 8.

Harlan JR, De Wet JJM. 1972. A Simplified Classification of Cultivated Sorghum1. *Crop Science* 12: 172–176.

Harris K, Subudhi PK, Borrell A, Jordan D, Rosenow D, Nguyen H, Klein P, Klein R, Mullet J. 2007. Sorghum stay-green QTL individually reduce post-flowering drought-induced leaf senescence. *Journal of Experimental Botany* 58: 327–338.

Haussmann B, Mahalakshmi V, Reddy B, Seetharama N, Hash C, Geiger H. 2002. QTL mapping of stay-green in two sorghum recombinant inbred populations. *Theoretical and Applied Genetics* 106: 133–142.

Hayes CM, Weers BD, Thakran M, Burow G, Xin Z, Emendack Y, Burke JJ, Rooney WL, Mullet JE. 2016. Discovery of a Dhurrin QTL in Sorghum: Co-localization of Dhurrin Biosynthesis and a Novel Stay-green QTL. *Crop Science* 56: 104–112.

Leiser WL, Rattunde HFW, Weltzien E, Cisse N, Abdou M, Diallo A, Tourè AO, Magalhaes JV, Haussmann BI. 2014. Two in one sweep: aluminum tolerance and grain yield in P-limited soils are associated to the same genomic region in West African Sorghum. *BMC Plant Biology* 14.

Li D, Dossa K, Zhang Y, Wei X, Wang L, Zhang Y, Liu A, Zhou R, Zhang X. 2018. GWAS Uncovers Differential Genetic Bases for Drought and Salt Tolerances in Sesame at the Germination Stage. *Genes* 9: 87.

Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang Z. 2012. GAPIT: genome association and prediction integrated tool. *Bioinformatics* 28: 2397–2399.

Mace E, Innes D, Hunt C, Wang X, Tao Y, Baxter J, Hassall M, Hathorn A, Jordan D. 2019. The Sorghum QTL Atlas: a powerful tool for trait dissection, comparative genomics and crop improvement. *Theoretical and Applied Genetics* 132: 751–766.

Mauboussin J-C, Gueye J, N'Diaye M. 1977. L'amélioration du Sorgho au Sénégal. *AGRONOMIE TROPICALE XXXll*: 8.

McCouch SR, Wright MH, Tung C-W, Maron LG, McNally KL, Fitzgerald M, Singh N, DeClerck G, Agosto-Perez F, Korniliev P, *et al.*2016. Open access resources for genome-wide association mapping in rice. *Nature Communications* 7: 1–14.

Mendiburu, F. (2009). *Agricolae: Statistical Procedures for Agricultural Research.* https://rdrr.io/cran/agricolae/man/agricolae-package.html

Mullet J, Morishige D, McCormick R, Truong S, Hilley J, McKinley B, Anderson R, Olson SN, Rooney W. 2014. Energy Sorghum--a genetic model for the design of C4 grass bioenergy crops. *Journal of Experimental Botany* 65: 3479–3489.

Mundia CW, Secchi S, Akamani K, Wang G. 2019. A Regional Comparison of Factors Affecting Global Sorghum Production: The Case of North America, Asia and Africa's Sahel. *Sustainability* 11: 2135.

Orr HA. 1998. The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. *Evolution; International Journal of Organic Evolution* 52: 935–949.

Poland J. 2015. Breeding-assisted genomics. *Current Opinion in Plant Biology* 24: 119–124.

Peterson, B. G., Carl, P., Boudt, K., Bennett, R., Ulrich, J., Eric Zivot, Lestel, M., Balkissoon, K., & Wuertz, D. (2014). *PerformanceAnalytics: Econometric Tools for Performance and Risk Analysis. Version 1.4.4000 from R-Forge*. https://rdrr.io/rforge/PerformanceAnalytics/.

Rao NN, Prasad K, Kumar PR, Vijayraghavan U. 2008. Distinct regulatory role for RFL, the rice LFY homolog, in determining flowering time and plant architecture. *Proceedings of the National Academy of Sciences* 105: 3646–3651.

R Core Team, R. C. (2016). *A language and environment for statistical computing. R Foundation for statistical computing, 2015; Vienna, Austria*.

Romero Navarro JA, Willcox M, Burgueño J, Romay C, Swarts K, Trachsel S, Preciado E, Terron A, Delgado HV, Vidal V, *et al.*2017. A study of allelic diversity underlying flowering-time adaptation in maize landraces. *Nature Genetics* 49: 476–480.

Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, Long Q, Nordborg M. 2012. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genetics* 44: 825–830.

Shin J-H, Blay S, McNeney B, Graham J. 2006. LDheatmap: An R Function for Graphical Display of Pairwise Linkage Disequilibria Between Single Nucleotide Polymorphisms | Shin | Journal of Statistical Software. 16.

Sinclair TR, Holbrook NM, Zwieniecki MA. 2005. Daily transpiration rates of woody species on drying soil. *Tree Physiology* 25: 1469–1472.

Tuinstra MR, Grote EM, Goldsbrough PB, Ejeta G. 1997. Genetic analysis of post-flowering drought tolerance and components of grain development in Sorghum bicolor (L.) Moench. *Molecular Breeding* 3: 439–448.

Wendorf F, Close AE, Schild R, Wasylikowa K, Housley RA, Harlan JR, Królik H. 1992. Saharan exploitation of plants 8,000 years BP. *Nature* 359: 721–724.

Xu W, Subudhi PK, Crasta OR, Rosenow DT, Mullet JE, Nguyen HT. 2000. Molecular mapping of QTLs conferring stay-green in grain sorghum (Sorghum bicolor L. Moench). *Genome* 43: 461–469.

Yano K, Yamamoto E, Aya K, Takeuchi H, Lo P, Hu L, Yamasaki M, Yoshida S, Kitano H, Hirano K, *et al.*2016. Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nature Genetics* 48: 927–934.

Yu J, Buckler ES. 2006. Genetic association mapping and genome organization of maize. *Current Opinion in Biotechnology* 17: 155–160.

Yuan Y, Xing H, Zeng W, Xu J, Mao L, Wang L, Feng W, Tao J, Wang H, Zhang H, *et al.*2019. Genome-wide association and differential expression analysis of salt tolerance in Gossypium hirsutum L at the germination stage. *BMC Plant Biology* 19: 394.

Zhao Y, Qiang C, Wang X, Chen Y, Deng J, Jiang C, Sun X, Chen H, Li J, Piao W, *et al.*2019. New alleles for chlorophyll content and stay-green traits revealed by a genome wide association study in rice ( Oryza sativa ). *Scientific Reports* 9: 2541.

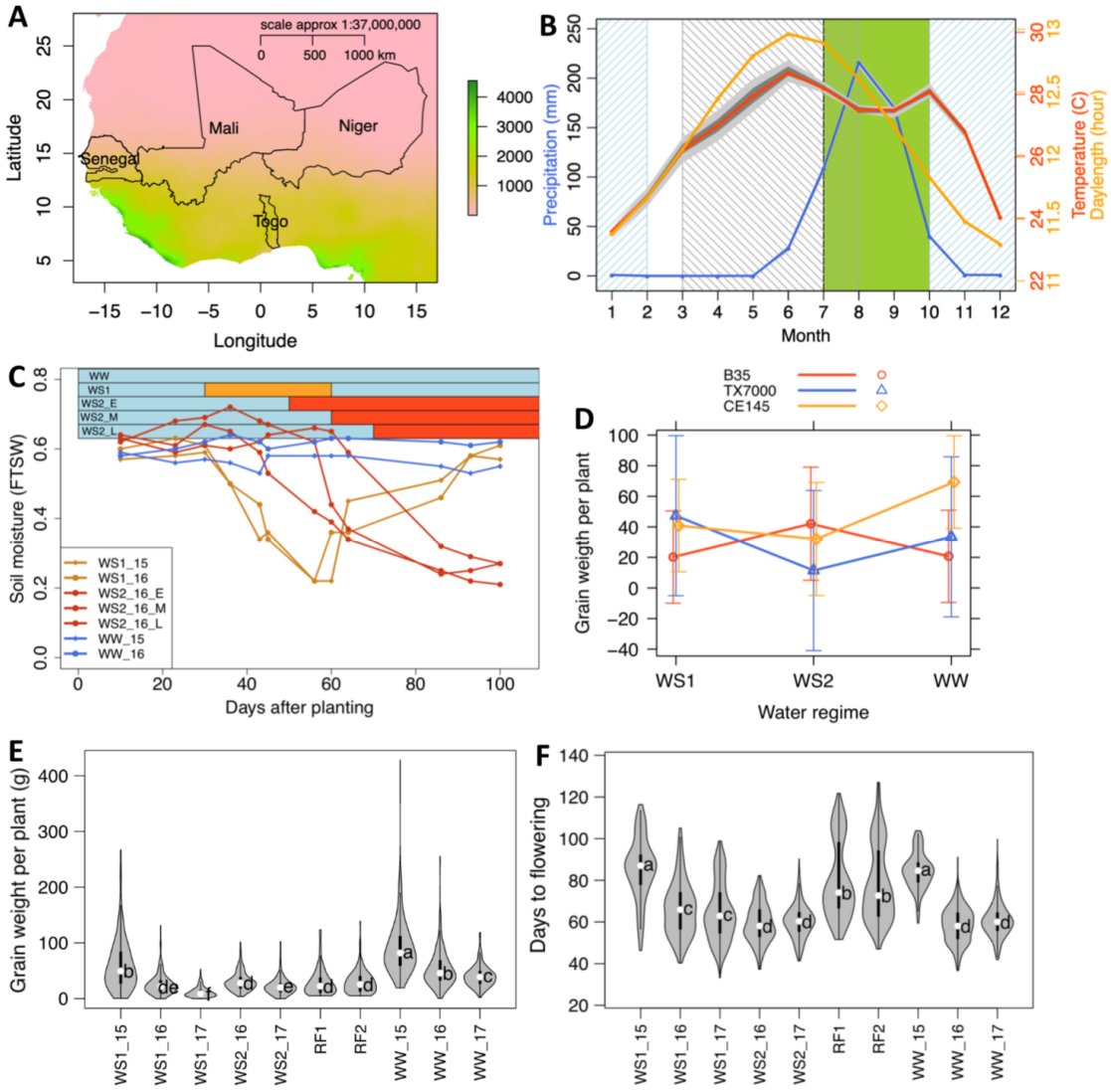



**Figure 4-1. Climatic variation and water deficit effect on managed drought environments.** (A) Precipitation gradient across sorghum accessions of origin in West Africa. The four countries of origin of accessions are represented in the West Africa map. (B) Average monthly precipitation, temperature, and daylength at the experimental station in Bambey, Senegal. (C) Fraction of transpirable soil water in well water environments (WW, blue lines), pre-flowering water stress environments (WS1, orange lines), and post-flowering water stress environments (WS2, red lines) during 2015 (line with diamond shape dots) and 2016 (line with close circle dots) off-seasons. The three lines of WS2 in 2016 represent three maturity groups (E, early maturity; M, medium maturity; L, late maturity). Horizontal bars indicate the water stress application periods for WS1 (orange) and WS2 (red) relative to WW (light blue). (D) Cross genotype x environment interaction of the pre-flowering (Tx7000, blue) and post-flowering

(B35, red) drought reference checks across WS1, WS2, and WW water regimes. The local drought tolerance check (CE145-266) is represented by the orange line. Effect of water deficit on (E) grain weight per plant and (F) days to flowering of accessions in each environment, including the two environments under rainfed conditions (RF1 and RF2).

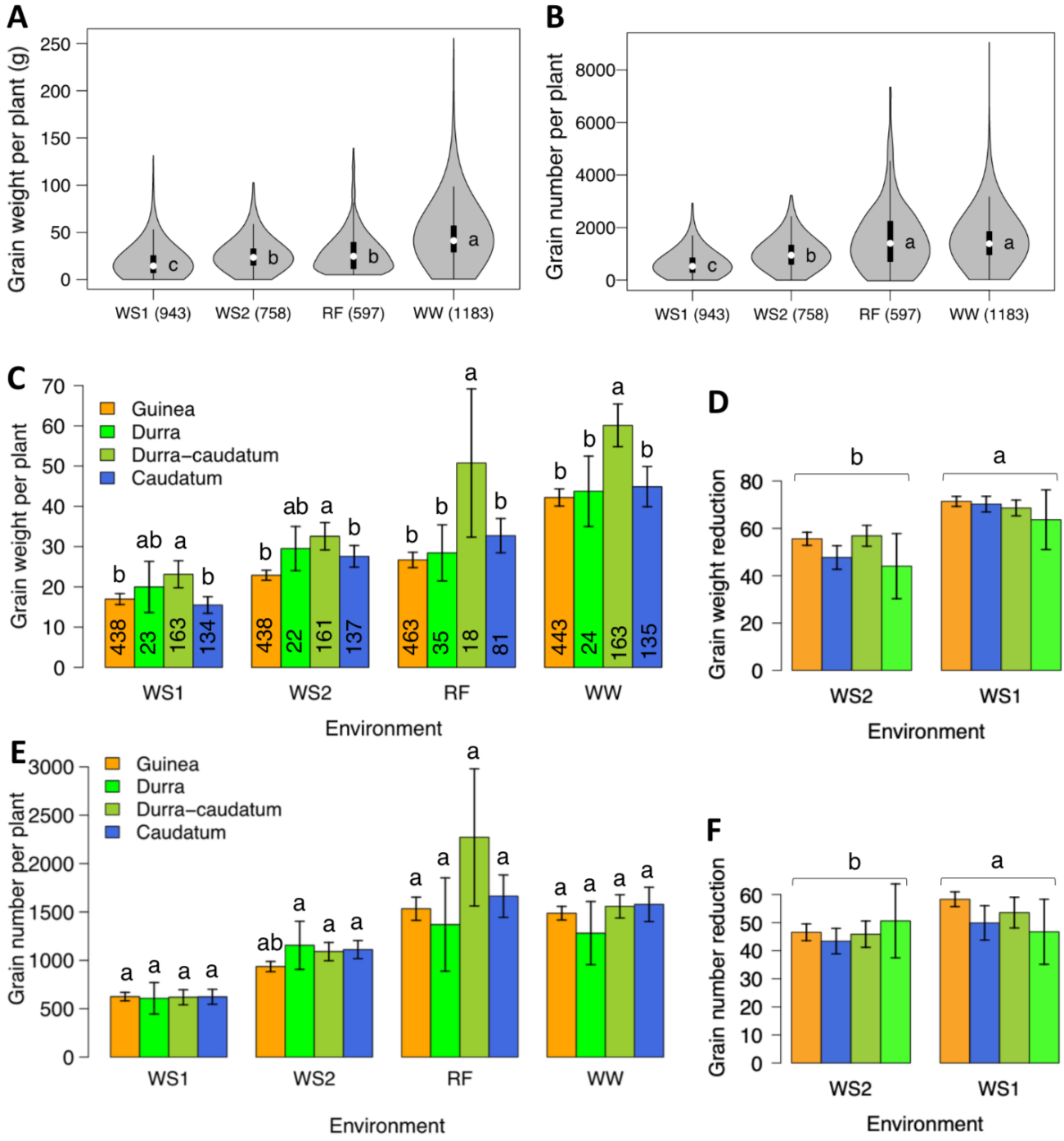
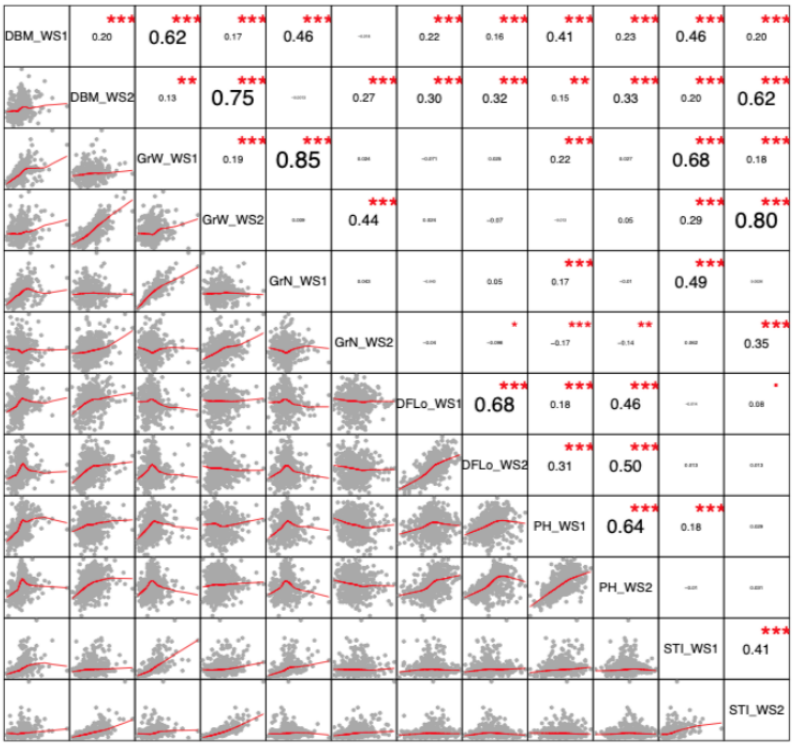



**Figure 4-2. Effect of water deficit on grain yield of accessions among water regimes.** Average values for (A) grain weight per plant and (B) grain number per plant in each water regime (WS1, pre-flowering water stress; WS2, post-flowering water stress; RF, rainfed conditions; WW, well water environments). The 2015 data was excluded. Letters within violin

plots indicate the Tuckey's HDS significance test. (C) Differences in grain weight among botanical types within each water regime. Digits within bar plots indicate the number of genotypes per botanical type in each water regime (two environments in each). (D) Percent reduction of grain weight among botanical types in stressed environments relative to control environments. (E) Differences in grain number among botanical types within each water regime. (F) Percent reduction of grain number among botanical types in stressed environments relative to control environments.

**A**


DBM_WS1
DBM_WS2
GrW_WS1
GrW_WS2
GrN_WS1
GrN_WS2
DFLo_WS1
DFLo_WS2
PH_WS1
PH_WS2
STI_WS1
STI_WS2

**B**

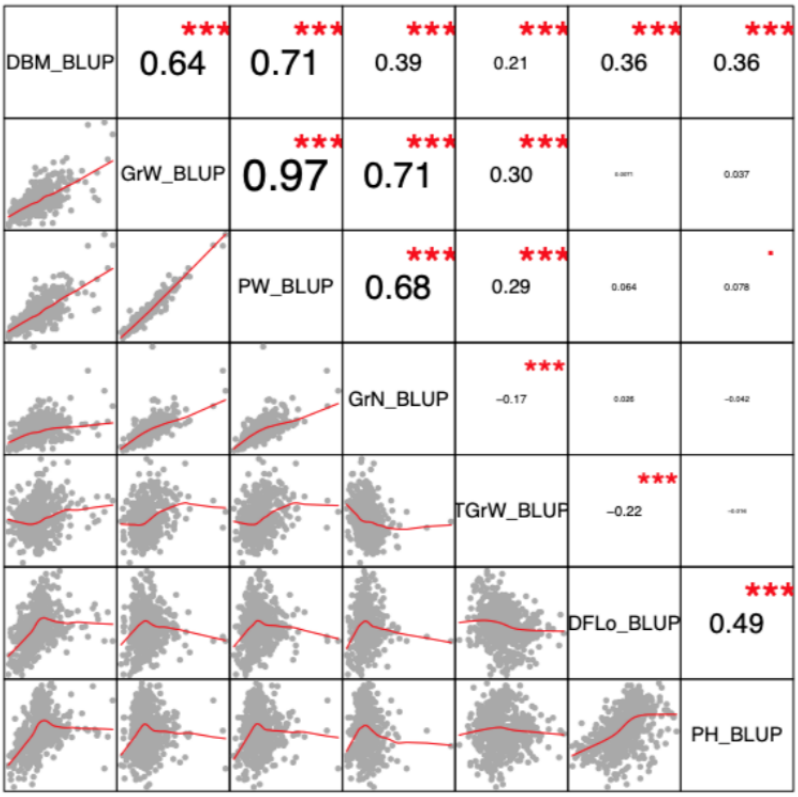

DBM_BLUP
GrW_BLUP
PW_BLUP
GrN_BLUP
TGrW_BLUP
DFLo_BLUP
PH_BLUP

**Figure 4-3. Phenotypic correlations of accessions.** (A) Correlations for yield components based on BLUP values in pre-flowering (WS1) and BLUP values in post-flowering (WS2) water stress environments. (B) Correlations for yield components based on BLUP values across all environments. DBM, above-ground dry biomass; GrW, grain weight per plant; PW, panicle weight per plant; GrN, grain number per plant; TGrW, thousand grain weight; DFLo, days to flowering; and PH, plant height.

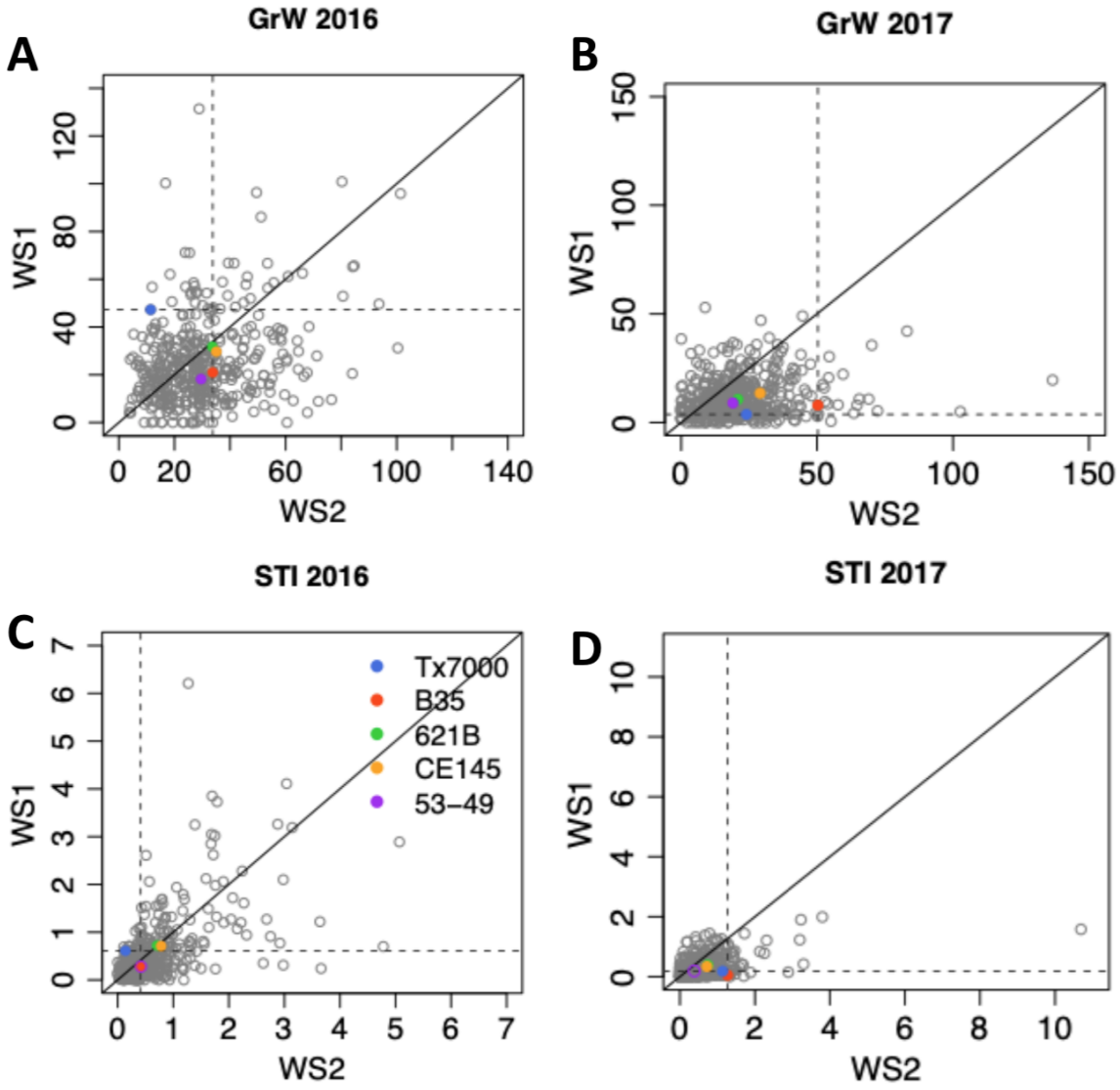



**Figure 4-4. Genotype performance in both pre- and post-flowering water stress.** (A) The 1:1 ratio correlation for grain weight per plant (GrW) and stress tolerance index (STI) for GrW of genotypes in pre-flowering (WS1) and post-flowering (WS2) water stress environments of 2016 and 2017. Color-coded dots indicate the pre-flowering (Tx7000) and post-flowering (B35) drought reference check lines, local drought tolerance check variety (CE145-266), and elite varieties (621B or Faourou and 53-49).

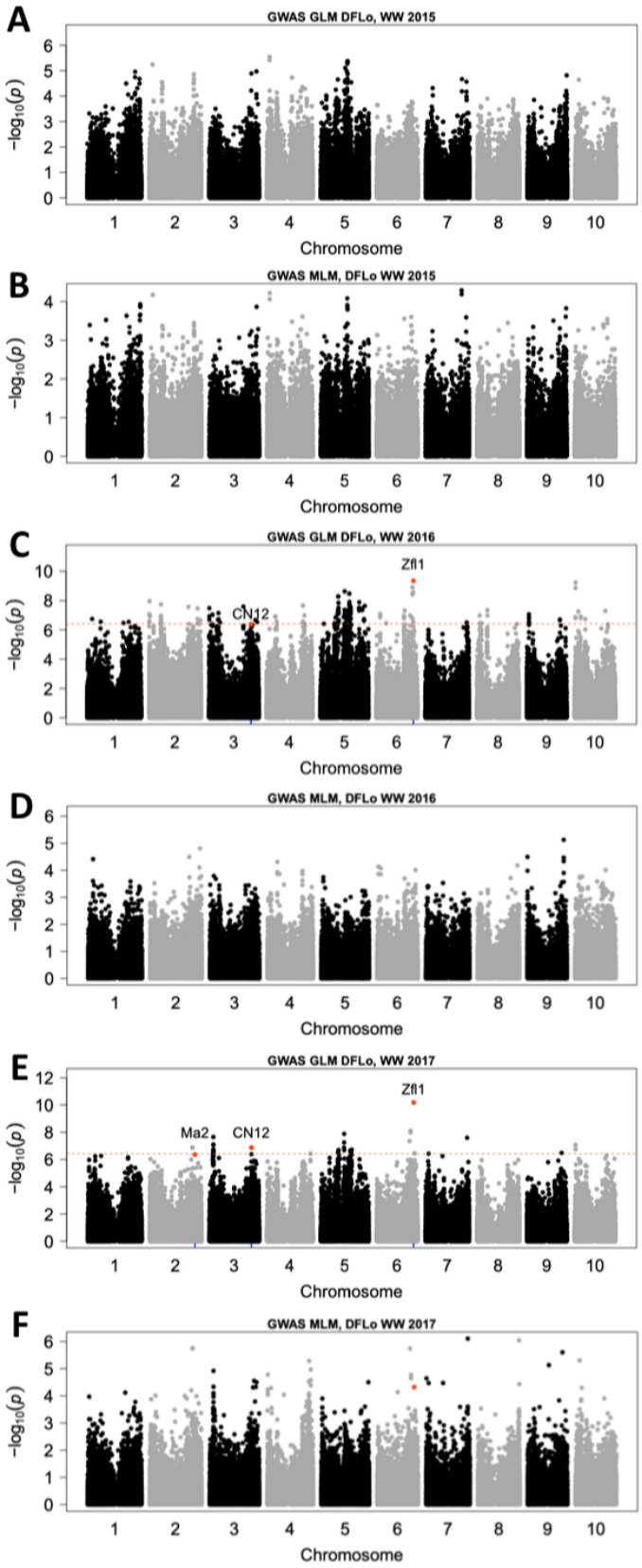

A
GWAS GLM DFLo, WW 2015
B
GWAS MLM, DFLo WW 2015
C
GWAS GLM DFLo, WW 2016
Zf1
CN12
D
GWAS MLM, DFLo WW 2016
E
GWAS GLM DFLo, WW 2017
Zf1
Ma2
CN12
F
GWAS MLM, DFLo WW 2017
−log10(p)
Chromosome

**Figure 4-5 GWAS for days to flowering (DFLo) under well-watered environments over three years**. Manhattan plots for days to flowering in 2015 using (A) general-linear model (GLM) with principal components and (B) mixed-linear model (MLM). Manhattan plots for days to flowering in 2016 using (C) GLM and (D) MLM. Manhattan plots for days to flowering in 2017 using (E) GLM and (F) MLM. Horizontal dashed line indicates the Bonferroni correction at 0.05. Red dots indicate peak SNPs colocalizing (based on 150 kb linkage disequilibrium decay rate) with flowering time candidate genes.

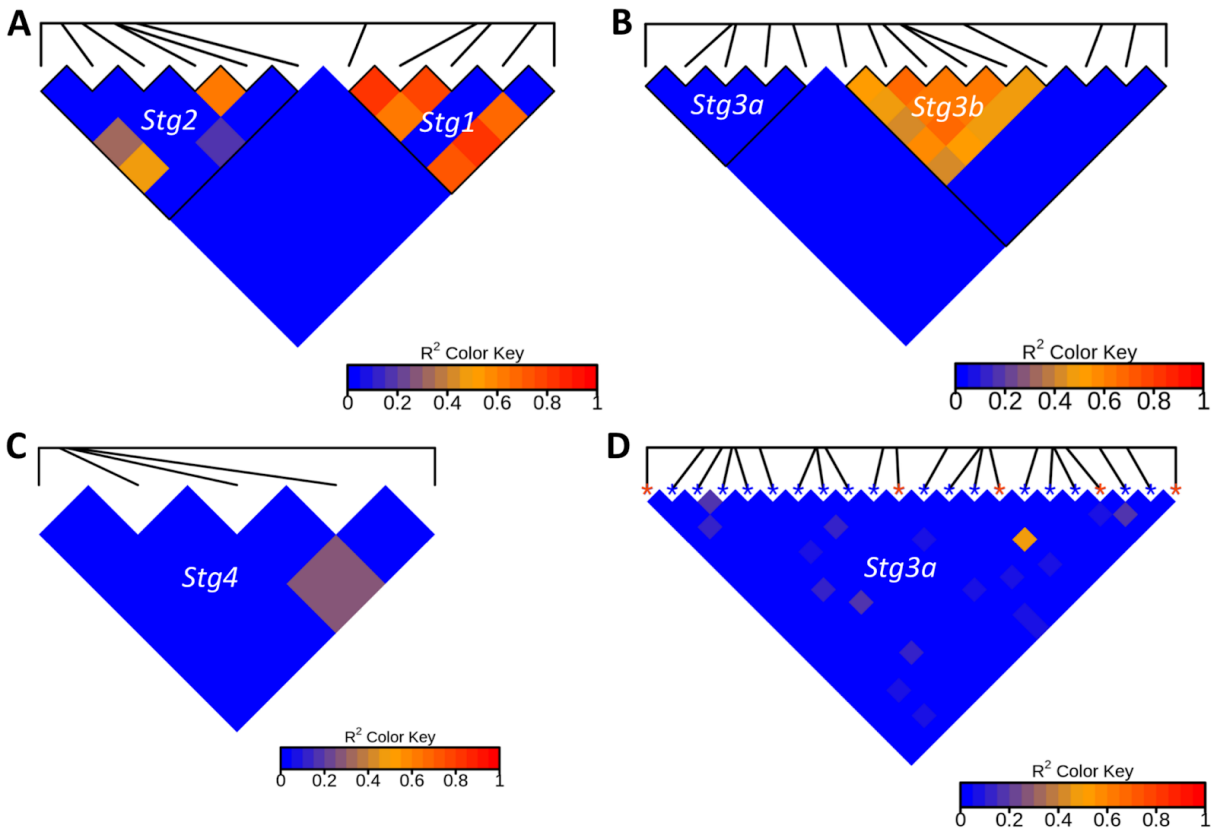



**Figure 4-6. Linkage disequilibrium heatmap for lead SNP associations at *Stg1–4* loci.** (A) Heatmap for lead SNPs at *Stg2* (left triangle) and *Stg1* (right triangle). (B) Heatmap for lead SNPs at *Stg3a* (left triangle) and *Stg3b* (right triangle). (C) Heatmap for lead SNPs at *Stg4*. (D) Linkage disequilibrium heatmap of lead SNP associations (red asterisks) and non-synonymous SNPs (blue asterisks) at *Stg3a* locus. Lead SNPs, S2_60973403 (second red asterisk from right) and S2_59237127 (second red asterisk from left) that are in moderate ($r^2 < 0.3$) and low ($r^2 < 0.1$) LD with a non-synonymous SNP, S2_61595689 (first blue asterisk from right).

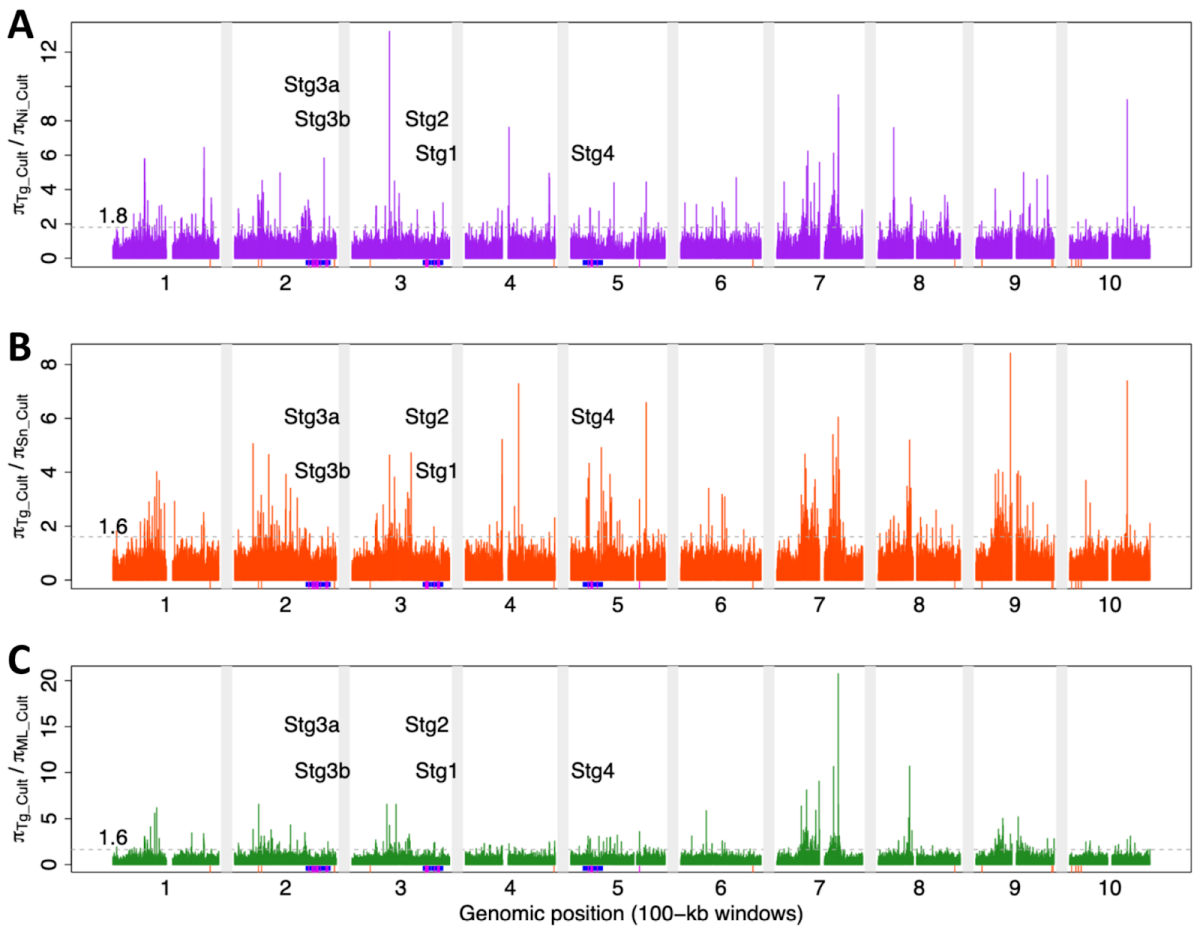



**Figure 4-7. Genomic selective sweeps to dry relative to humid environments**. Reduction of pairwise nucleotide diversity (*pi*) around drought-yield QTLs in the genome of (A) Niger, (B) Senegal, (C) Mali accessions relative to Togo accessions of the WASAP. Reduction of nucleotide diversity was calculated based on 100-kb sliding windows. Dashed horizontal lines indicate the threshold for the top 5% signatures of selection outliers. Dashed vertical lines indicate the genomic position of the colocalized Stay-green QTLs (*Stg1–4*) with signatures of selection outliers. The Rug-plots in red indicate the genomic position of the pleiotropic lead SNPs associated with drought response variables. The Rug-plots in magenta indicate the genomic position of lead SNP within *Stg1–4*.

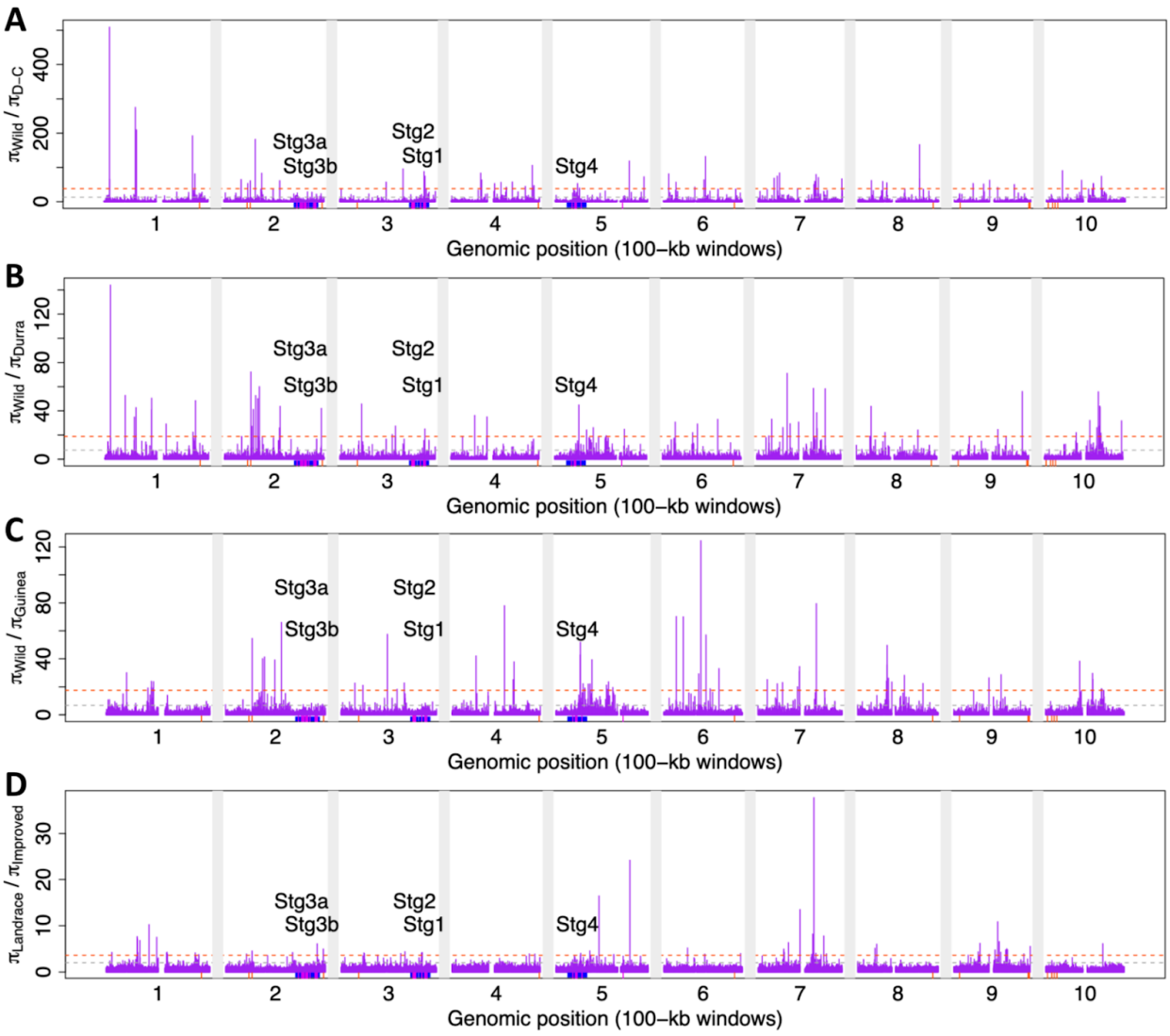



**Figure 4-8. Selective sweeps in drought tolerance loci for domestication and improvement.** Signatures of selection colocalizations with drought response QTLs in (A) durra-caudatum (D-C) landraces, (B) durra landraces, (C) guinea landraces, and (D) improved lines. The reduction of nucleotide diversity was calculated based on 100-kb sliding windows. Dashed horizontal lines indicate the threshold for the top 5% signatures of selection outliers. Blue segments indicate the genomic position of the colocalized stay-green QTLs (*Stg1–4*) surrounding signatures of selection outliers. The red Rug-plots indicate the genomic position of the pleiotropic lead SNPs associated with drought response variables.

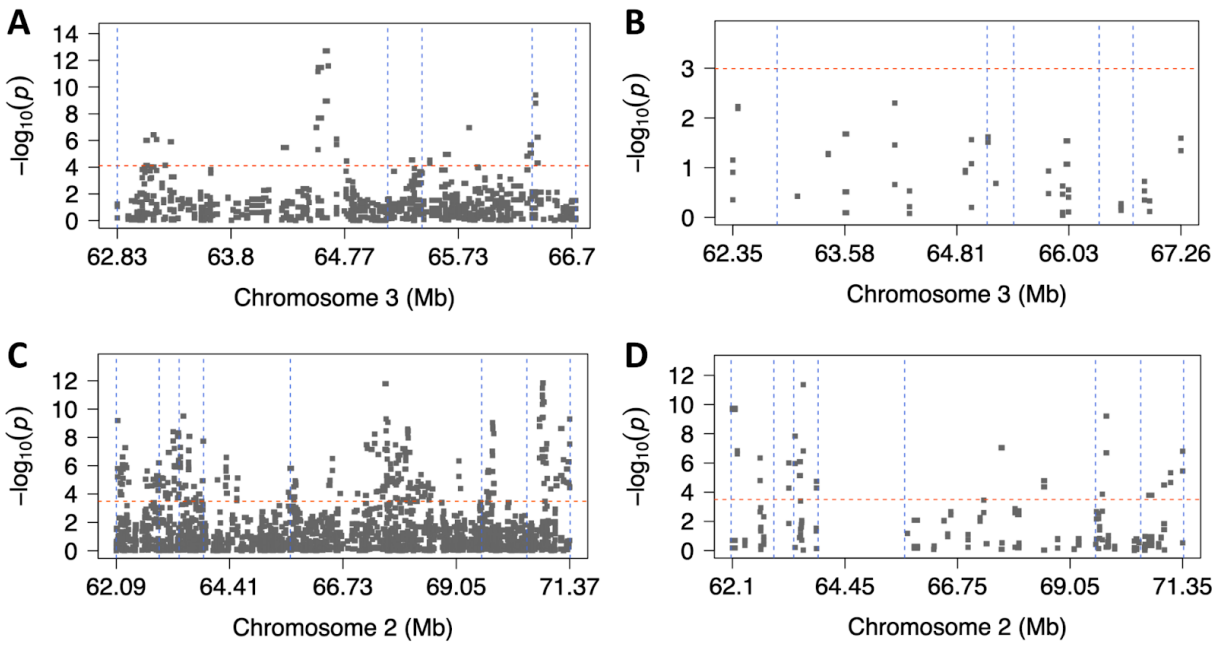



**Figure 4-9. Haplotype-based associations of drought tolerance quantitative trait loci.** Regional Manhattan plot for haplotype blocks estimated based on (A) all SNPs and (B) non-synonymous SNPs within *Stg1*. Regional Manhattan plot for haplotype blocks estimated based on (C) all SNPs and (D) non-synonymous SNPs within *Stg3b*. Horizontal dashed lines indicate the Bonferroni correction at 0.05. Vertical dashed lines indicate the position of lead SNP associations that are colocalized with *Stg1* or *Stg3b*.

**Table 4-1. Descriptive statistics, variance components, and broad-sense heritability ($H^2$) of yield components across all environments.**

| Trait | G (%) | E (%) | G x E (%) | $H^2$ | Range | Mean ± SD | CV (%) |
|---|---|---|---|---|---|---|---|
| DBM (g) | 3 | 73 | 8 | 0.65 | 6–1311 | 138 ± 144 | 105 |
| GrW (g) | 4 | 51 | 31 | 0.71 | 0–428 | 37 ± 34 | 94 |
| GrN | 3 | 50 | 24 | 0.53 | 0–17780 | 1315 ± 1210 | 92 |
| TGrW (g) | 35 | 9 | 28 | 0.91 | 0–56 | 28 ± 8 | 29 |
| DFLo | 24 | 56 | 9 | 0.95 | 33–116 | 64 ± 14 | 22 |
| PH (cm) | 40 | 31 | 13 | 0.95 | 40–356 | 182 ± 54 | 30 |

DBM, dry biomass per plant; GrW, grain weight per plant; GrN, grain number per plant; TGrW, thousand grain weight; DFLo, days to flowering; PH, plant height; SD, standard deviation; CV, coefficient of variation; G, genotype and E, environment variances; $H^2$, broad-sense heritability.

**Table 4-2. GWAS pleiotropic lead SNPs for reduction of yield components and stress tolerance index for grain weight (STI) in independent and across water stress environments.**

| Lead SNP [a] | *P* value | MAF | Effect | Trait | Env. | Model | *BLUP* $R^2$ [b] | *BLUP P* value [c] |
|---|---|---|---|---|---|---|---|---|
| S1_74186408 | $<10^{-7}$ | 0.03 | -106 | RPW | WS2 2017 | MLM | 0.11 | $<10^{-8}$ |
| | $<10^{-11}$ | | -116 | | WS1 2017 | MLM | | |
| S2_18195896 | $<10^{-9}$ | 0.02 | 0.32 | STI | WS1 2017 | GLM, MLM | 0.22 | $<10^{-16}$ |
| | $<10^{-17}$ | | 1.07 | | WS2 2017 | GLM, MLM | | |
| S2_20558788 | $<10^{-19}$ | 0.03 | 1.50 | STI | WS2 2017 | GLM, MLM | 0.18 | $<10^{-15}$ |
| | $<10^{-7}$ | | 0.33 | | WS1 2017 | GLM, MLM | | |
| S2_76213690 | $<10^{-13}$ | 0.04 | -0.73 | STI | WS2 2017 | GLM, MLM | 0.16 | $<10^{-13}$ |
| | $<10^{-9}$ | | -0.25 | | WS1 2017 | GLM, MLM | | |
| S3_13763609 | $<10^{-7}$ | 0.02 | -110 | RPW | WS2 2017 | MLM | 0.12 | $<10^{-9}$ |
| | $<10^{-11}$ | | -145 | | WS1 2017 | MLM | | |
| S3_56094063 | $<10^{-19}$ | 0.02 | 1.34 | STI | WS2 2017 | GLM, MLM | 0.19 | $<10^{-16}$ |
| | $<10^{-8}$ | | 0.38 | | WS1 2017 | GLM, MLM | | |
| S4_67777846 | $<10^{-17}$ | 0.03 | -1.20 | STI | WS2 2017 | GLM, MLM | 0.25 | $<10^{-6}$ |
| | $<10^{-8}$ | | -1.19 | | WS1 2016 | GLM, MLM | | |
| | $<10^{-8}$ | | -0.35 | | WS1 2017 | GLM, MLM | | |
| S6_55048997 | $<10^{-7}$ | 0.17 | -0.30 | STI | WS2 2016 | GLM | 0.16 | $<10^{-13}$ |
| | $<10^{-9}$ | | -0.38 | | WS1 2016 | GLM | | |
| S8_58355080 | $<10^{-19}$ | 0.02 | 1.33 | STI | WS2 2017 | GLM, MLM | 0.18 | $<10^{-16}$ |
| | $<10^{-10}$ | | 0.42 | | WS1 2017 | GLM, MLM | | |
| S9_4530433 | $<10^{-14}$ | 0.02 | 0.99 | STI | WS2 2017 | GLM | 0.17 | $<10^{-14}$ |
| | $<10^{-7}$ | | 0.31 | | WS1 2017 | GLM | | |

**Table 4-2. Continue**

| Lead SNP [a] | *P* value | MAF | Effect | Trait | Env. | Model | BLUP $R^2$ [b] | BLUP *P* value [c] |
|---|---|---|---|---|---|---|---|---|
| S9_57781496 | $<10^{-16}$ | 0.02 | -0.98 | STI | WS2 2017 | GLM, MLM | 0.20 | $<10^{-16}$ |
| | $<10^{-10}$ | | -0.31 | | WS1 2017 | GLM, MLM | | |
| S9_58763841 | $<10^{-16}$ | 0.02 | 1.14 | STI | WS2 2017 | GLM, MLM | 0.18 | $<10^{-15}$ |
| | $<10^{-7}$ | | 0.31 | | WS1 2017 | GLM | | |
| S10_1402513 | $<10^{-8}$ | 0.14 | 0.59 | STI | WS1 2016 | GLM, MLM | 0.14 | $<10^{-15}$ |
| | $<10^{-11}$ | | 0.68 | | WS2 2016 | GLM, MLM | | |
| S10_4711152 | $<10^{-6}$ | 0.04 | -67 | RPW | WS1 2015 | MLM | 0.11 | $<10^{-8}$ |
| | $<10^{-6}$ | | -69 | RGrN | WS1 2015 | MLM | | |
| S10_6619068 | $<10^{-13}$ | 0.03 | 0.90 | STI | WS2 2017 | GLM | 0.16 | $<10^{-13}$ |
| | $<10^{-8}$ | | 0.31 | | WS1 2017 | GLM, MLM | | |
| S10_8716926 | $<10^{-7}$ | 0.02 | -0.36 | STI | WS1 2017 | GLM | 0.18 | $<10^{-15}$ |
| | $<10^{-16}$ | | -1.47 | | WS2 2017 | GLM, MLM | | |

[a] Digit before and after underscore indicates chromosome number and SNP position on the genome, respectively; [b] proportion of phenotypic variation explained based on BLUPs across drought water stress environments and ADMIXTURE ancestry memberships at K = 8 were used as fixed effect covariate; [c] significance of proportion of phenotypic variation explained; MAF, minor allele frequency; GLM, general linear model; MLM, mixed-linear model; Env, water stress environments.

**Table 4-3. GWAS lead SNPs within *Stg1–4* loci for reduction of yield components and stress tolerance index for grain weight (STI) in independent and across water stress environments.**

| Lead SNP [a] | *P* value | MAF | Effect | Trait | Env. | Model | BLUP $R^2$ [b] | Locus |
|---|---|---|---|---|---|---|---|---|
| S2_56682379 | $<10^{-7}$ | 0.02 | -0.6 | STI | WS1 2016 | GLM, MLM | 0.16 | *Stg3a* |
| S2_59129283 | $<10^{-10}$ | 0.03 | 0.9 | STI | WS2 2017 | GLM, MLM | 0.13 | |
| S2_59237127 | $<10^{-7}$ | 0.03 | 0.6 | STI | WS1 2016 | GLM, MLM | 0.13 | |
| S2_60191986 | $<10^{-7}$ | 0.02 | 1.3 | STI | WS1 2015 | GLM, MLM | 0.12 | |
| S2_60849014 | $<10^{-8}$ | 0.02 | -95 | RPW | WS1 2017 | MLM | 0.12 | |
| S2_62095163 | $<10^{-18}$ | 0.02 | 1.5 | STI | WS2 2017 | GLM, MLM | 0.21 | *Stg2b* |
| S2_62973945 | $<10^{-16}$ | 0.02 | 1.2 | STI | WS2 2017 | GLM, MLM | 0.18 | |
| S2_63381610 | $<10^{-17}$ | 0.02 | 1.2 | STI | WS2 2017 | GLM, MLM | 0.20 | |
| S2_63881780 | $<10^{-13}$ | 0.02 | 1.1 | STI | WS2 2017 | GLM, MLM | 0.22 | |
| S2_65658140 | $<10^{-13}$ | 0.02 | 1 | STI | WS2 2017 | GLM, MLM | 0.15 | |
| S2_69575903 | $<10^{-5}$ | 0.17 | -38 | RPW | WS1 2017 | MLM | 0.11 | |
| S2_70503173 | $<10^{-8}$ | 0.04 | -0.6 | STI | WS2 2016 | GLM, MLM | 0.12 | |
| S2_71386056 | $<10^{-5}$ | 0.04 | 12 | RDBM | WS1 2016 | MLM | 0.12 | |
| S3_56094063 | $<10^{-19}$ | 0.02 | 1.3 | STI | WS1, WS2 2017 | GLM, MLM | 0.19 | *Stg2* |
| S3_56515341 | $<10^{-7}$ | 0.04 | 17 | RGrN | WS1 2017 | GLM, MLM | 0.12 | |
| S3_56946323 | $<10^{-5}$ | 0.07 | 9 | RDBM | WS1 2016 | MLM | 0.11 | |
| S3_57614567 | $<10^{-12}$ | 0.02 | 0.8 | STI | WS2 2017 | GLM | 0.16 | |

**Table 4-3. Continue**.

| Lead SNP [a] | *P* value | MAF | Effect | Trait | Env. | Model | BLUP $R^2$ [b] | Locus |
|---|---|---|---|---|---|---|---|---|
| S3_57615696 | $<10^{-7}$ | 0.02 | -0.3 | STI | WS1 2017 | GLM | 0.16 | |
| S3_58067325 | $<10^{-7}$ | 0.06 | -1 | STI | WS1 2015 | GLM, MLM | 0.11 | |
| S3_62836558 | $<10^{-8}$ | 0.02 | -72 | RPW | WS1 2017 | GLM | 0.12 | *Stg1* |
| S3_65137990 | $<10^{-10}$ | 0.02 | -73 | RPW | WS1 2017 | GLM, MLM | 0.11 | |
| S3_65430305 | $<10^{-7}$ | 0.02 | 96 | RPW | WS2 2017 | MLM | 0.11 | |
| S3_66366589 | $<10^{-8}$ | 0.03 | -0.7 | STI | WS1 2016 | GLM, MLM | 0.15 | |
| S3_66738018 | $<10^{-11}$ | 0.03 | 72 | RPW | WS1 2017 | GLM, MLM | 0.11 | |
| S5_13190947 | $<10^{-11}$ | 0.03 | -0.7 | STI | WS2 2017 | GLM, MLM | 0.16 | *Stg4* |
| S5_15215761 | $<10^{-8}$ | 0.03 | -43 | RPW | WS1 2017 | GLM, MLM | 0.11 | |
| S5_15916423 | $<10^{-8}$ | 0.03 | -71 | RPW | WS1 2017 | GLM, MLM | 0.11 | |
| S5_16480120 | $<10^{-10}$ | 0.03 | -1.2 | STI | WS1 2015 | GLM, MLM | 0.16 | |
| S5_20251208 | $<10^{-14}$ | 0.02 | -1 | STI | WS2 2017 | GLM, MLM | 0.19 | |
| S5_52255304 | $<10^{-13}$ | 0.03 | -1.5 | STI | WS2 2017 | GLM, MLM | 0.14 | |

[a] Digit before and after underscore indicates chromosome number and SNP position on the genome, respectively; [b] proportion of phenotypic variation explained based on BLUPs across water stress environments and ADMIXTURE ancestry memberships at K = 8 were used as fixed effect covariate; MAF, minor allele frequency; GLM, general linear model; MLM, mixed-linear model; Env, water stress environments.

**Table 4-4. Pairwise-wide nucleotide diversity and frequency of common allele of lead SNP associations under positive selection.**

| | | | Nucleotide diversity ratio | | | Common allele frequency | | |
|---|---|---|---|---|---|---|---|---|
| Chr | Locus | Lead SNP | πW/πDC | πW/πD | πW/πG | DC | Durra | Guinea |
| 1 | – | 74186408 | 18.2 | **3** | – | 0.5 | 0.5 | 0.46 |
| 2 | – | 18195896 | **2.4** | 9 | – | 0.49 | 0.5 | 0.47 |
| 2 | – | 20558788 | 61.3 | 72.2 | 54.6 | 0.49 | 0.5 | 0.45 |
| 2 | *Stg3b* | 71386056 | **2.1** | 7.8 | – | 0.5 | 0.5 | 0.44 |
| 2 | – | 76213690 | 13.4 | **3.3** | – | 0.45 | 0.26 | – |
| 3 | *Stg1* | 62836558 | **2.8** | **2.2** | 7.5 | 0.5 | 0.5 | 0.45 |
| 3 | *Stg1* | 65137990 | 15.1 | **1.7** | – | 0.5 | 0.5 | 0.46 |
| 3 | *Stg1* | 65430305 | 88.2 | 15.5 | 8.4 | 0.5 | 0.5 | 0.46 |
| 3 | *Stg1* | 66366589 | 74.6 | 25 | – | 0.5 | 0.46 | 0.46 |
| 5 | *Stg4* | 15215761 | 32 | **6.8** | 10.1 | 0.5 | 0.48 | 0.48 |
| 5 | *Stg4* | 16480120 | 27 | 13 | – | 0.45 | 0.41 | 0.48 |
| 5 | *Stg4* | 20251208 | **1** | **1.5** | 8.7 | 0.45 | 0.46 | 0.5 |

Bold numbers are not among the 95 percentile selective sweep outliers.
Chr, chromosome; DC, durra-caudatum landraces; D, durra landraces; G, guinea landraces; W, wild relative sorghum

**Table 4-5. The top two significantly associated haplotypes at *Stg1* quantitative trait locus.**

| | | | | | Haplotype allele frequency | | | |
|---|---|---|---|---|---|---|---|---|
| **Left SNP** | **Right SNP** | **Haplotype** | ***P*-value[a]** | ***beta*[b] (%)** | **D-C** | **Durra** | **Guinea** | **Improved** |
| S3_64606653 | S3_64623404 | AGC | $<10^{-13}$ | 3.5 | 0.63 | 0 | 0 | 0 |
| | | CAC | 0.62 | 0.3 | 0 | 0.13 | 0.05 | 0.21 |
| | | CAT | $<10^{-9}$ | -2.3 | 0.33 | 0.85 | 0.95 | 0.79 |
| | | CGT | 0.75 | -3.1 | 0 | 0.02 | 0 | 0 |
| S3_66396518 | S3_66396621 | TA | $<10^{-10}$ | 2.1 | 0.76 | 0.61 | 0.15 | 0.27 |
| | | CT | $<10^{-9}$ | -2.0 | 0.23 | 0.39 | 0.83 | 0.73 |
| | | TT | 0.04 | -3.5 | 0 | 0 | 0.01 | 0 |

[a] The *p*-values for haplotype-specific tests of association with stress tolerance index for grain weight across drought stress environments in the whole WASAP panel; [b] *beta*, regression coefficient of haplotype alleles in the WASAP; D-C, durra caudatum landraces; improved, improved varieties.

**Table 4-6. The top two most significant haplotype associations and haplotypes that overlap with lead SNPs at *Stg3b* quantitative trait locus.**

| | | | | | Haplotype allele frequency | | | |
|---|---|---|---|---|---|---|---|---|
| **Left SNP** | **Right SNP** | **Haplotype[a]** | ***P*-value[b]** | **Beta[c] (%)** | **D-C** | **Durra** | **Guinea** | **Improved** |
| S2_70827637 | S2_70834834 | TGGATGA | $<10^{-12}$ | 3.1 | 0.67 | 0.07 | 0 | 0.03 |
| | | TCGACTT | $<10^{-10}$ | -2.0 | 0.21 | 0.46 | 0.81 | 0.83 |
| | | TCGATGA | 0.02 | 1.3 | 0.05 | 0.26 | 0.09 | 0.05 |
| | | GGTGTGA | 0.04 | 1.7 | 0.04 | 0.13 | 0 | 0.03 |
| | | TCGACGA | 0.5 | -1.3 | 0.02 | 0.02 | 0.01 | 0 |
| | | TGTGTGA | $<10^{-3}$ | -2.1 | 0 | 0.04 | 0.05 | 0.03 |
| | | GGGGTGA | 0.40 | -8.1 | 0 | 0.02 | 0 | 0 |
| S2_67597045 | S2_67619513 | GTACT | $<10^{-12}$ | 3.1 | 0.67 | 0 | 0.02 | 0.03 |
| | | CCTTG | $<10^{-8}$ | -1.8 | 0.09 | 0.22 | 0.67 | 0.03 |
| | | GCTCT | 0.05 | 1.1 | 0.04 | 0.26 | 0.08 | 0.05 |
| | | GCTTT | 0.61 | 0.5 | 0.03 | 0.22 | 0 | 0 |
| | | GCTTG | 0.07 | -0.8 | 0.13 | 0.30 | 0.21 | 0.29 |
| S2_62962601 | S2_62974532 | GTACTCAGGGCGG | $<10^{-7}$ | 6.8 | 0.02 | 0.04 | 0 | 0.02 |
| | | TTATCTCCCGCGG | $<10^{-5}$ | -1.5 | 0.10 | 0.09 | 0.54 | 0.24 |
| | | GGGCTCAGCATCA | $<10^{-3}$ | 1.7 | 0.37 | 0 | 0.05 | 0 |
| | | GGGCTCAGCATCG | 0.94 | -0.1 | 0 | 0 | 0.01 | 0.04 |
| | | GGGCTCAGCGCGG | 0.23 | 0.4 | 0.39 | 0.39 | 0.25 | 0.64 |
| | | GTACTCAGCGCGG | 0.99 | 0 | 0.05 | 0.44 | 0.07 | 0.03 |
| S2_63870593 | S2_63881780 | CA | $<10^{-8}$ | 7.7 | 0.02 | 0.09 | 0 | 0.02 |
| | | CT | $<10^{-2}$ | -0.6 | 0.16 | 0.48 | 0.38 | 0.57 |
| | | GT | 0.75 | 0.1 | 0.83 | 0.44 | 0.61 | 0.42 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| S2_65651910 | S2_65679884 | CGTTGGCACA | $<10^{-6}$ | 6.5 | 0.02 | 0.04 | 0 | 0.02 |
| | | AACTGGCGTA | $<10^{-4}$ | 1.5 | 0.58 | 0.11 | 0.14 | 0.14 |
| | | CACTGCAATG | $<10^{-3}$ | -1.2 | 0.07 | 0.09 | 0.41 | 0.10 |
| | | CACCGGCATA | 0.01 | -1.3 | 0 | 0 | 0.20 | 0.14 |
| | | CACTGGCATA | 0.31 | -0.8 | 0.06 | 0.17 | 0.03 | 0.03 |
| | | AACTGGCATA | 0.46 | -0.9 | 0.03 | 0 | 0 | 0 |
| | | CGCTAGCACA | 0.90 | 0.1 | 0.01 | 0.24 | 0.03 | 0.02 |
| | | CACTGGCACA | 0.14 | 0.6 | 0.15 | 0.28 | 0.12 | 0.51 |
| | | CACTGGCGTA | 0.33 | -3.7 | 0.02 | 0 | 0 | 0.02 |
| S2_69575903 | S2_69575906 | GA | $<10^{-4}$ | 2.1 | 0.05 | 0.07 | 0.05 | 0.03 |
| | | TG | $<10^{-2}$ | -0.8 | 0.15 | 0.25 | 0.35 | 0.03 |
| | | GG | 0.87 | -0.1 | 0.80 | 0.68 | 0.60 | 0.94 |

[a] The most significantly associated haplotype alleles of a given haplotype block only are included in this table; [b] The *p*-values for haplotype-specific tests of association with stress tolerance index for grain weight across drought stress environments in the whole WASAP panel; [c] Beta, percent regression coefficient in the WASAP; D-C, durra caudatum landraces; improved, improved varieties.

# Supplemental Materials Chapter 4

This section includes the supplemental figures and tables for the chapter 4

Genome-Wide Association Studies of Drought Tolerance in West African Sorghum

**Table C- 1. Summary statistics and variance components of check lines in each water regime.**

| Trait | WR | G (%) | E (%) | G x E (%) | Range | Mean ± SD | CV (%) |
|---|---|---|---|---|---|---|---|
| GrW (g) | Hiv | 34 | 52 | 1 | 16–47 | 31 ±10 | 33 |
| | WS1 | 31 | 34 | 9 | 3–47 | 18 ± 11 | 62 |
| | WS2 | 0 | 0 | 41 | 11–50 | 29 ± 9 | 31 |
| | WW | 32 | 0 | 50 | 9–71 | 41 ± 17 | 43 |
| DBM (g) | Hiv | 9 | 87 | 0 | 93–218 | 164 ± 42 | 26 |
| | WS1 | 12 | 67 | 7 | 16–160 | 65 ± 35 | 53 |
| | WS2 | 0 | 26 | 43 | 41–114 | 77 ± 22 | 28 |
| | WW | 7 | 66 | 7 | 35–193 | 102 ± 54 | 53 |
| GrN | Hiv | 25 | 46 | 7 | 812–3077 | 1995 ± 716 | 36 |
| | WS1 | 28 | 27 | 18 | 175–1645 | 738 ± 402 | 55 |
| | WS2 | 25 | 0 | 52 | 506–2071 | 1378 ± 383 | 28 |
| | WW | 47 | 0 | 26 | 351–2496 | 1525 ± 648 | 42 |
| TGrW (g) | Hiv | 40 | 51 | 0 | 10–16 | 14 ± 2 | 14 |
| | WS1 | 0 | 41 | 29 | 18–35 | 25 ± 5 | 21 |
| | WS2 | 26 | 0 | 45 | 15–31 | 21 ± 4 | 20 |
| | WW | 12 | 6 | 33 | 20–36 | 26 ± 4 | 14 |
| DFLo (days) | Hiv | 36 | 43 | 2 | 62–74 | 65 ± 4 | 6 |
| | WS1 | 51 | 0 | 25 | 53–88 | 65 ± 11 | 17 |
| | WS2 | 66 | 0 | 11 | 50–77 | 58 ± 6 | 10 |
| | WW | 80 | 0 | 4 | 50–75 | 59± 7 | 12 |

DBM, dry biomass per plant; GrW, grain weight per plant; GrN, grain number per plant; TGrW, thousand grain weight; DFLo, days to flowering; SD, standard deviation; CV, coefficient of variation; G, genotype and E, environment variances; WR, water regime.

# Chapter 5 - Genomics-Enabled Breeding for Crop Improvement in West Africa

## Genomics-Enabled Breeding in Classical Breeding Programs

My vision is to live in a world where smallholder farmers in semi-arid regions can produce sustainably sufficient food to improve their living standards regardless of climate changes. To contribute to this vision, my mission is to understand the genetic architecture of adaptive traits across the West African region to establish an efficient molecular breeding platform. One of the breeding priorities of Breeding programs in West Africa is to develop locally adapted varieties with yield advantages over existing varieties and photoperiodic flowering that fits the different agro-ecological regions. However, field phenotyping for optimal selection is difficult or even impossible in early generations, requiring multi-environment trials. Breeding programs are small with limited resources to handle multiple field trials across diverse environments, particularly for complex traits such as drought tolerance and photoperiodic flowering. The development of adaptive traits-associated markers can contribute to rapidly develop varieties for short-term and long-term delivery to growers.

## Marker-Assisted Backcrossing for Drought-Yield Improvement

My first research goal is to develop high-throughput breeder-friendly markers that predict drought tolerance, photoperiodic flowering, and panicle architecture to rapidly introgress favorable alleles from donor lines into locally preferred varieties. Marker-assisted backcrossing (MABC) is suitable for the introgression of a few large-effect QTLs (Varshney *et al.*, 2013). Toward this goal, my favored hypothesis regarding drought tolerance was that the genetic architecture of grain yield under drought in West African sorghum is oligogenic, with the contribution of large-effect drought tolerance alleles. Studies have demonstrated that genetic variation controls drought tolerance in sorghum (Tuinstra *et al.*, 1996, 1997; Xu *et al.*, 2000; Kebede *et al.*, 2001; Haussmann *et al.*, 2002; Harris *et al.*, 2007; Borrell *et al.*, 2014; Hayes *et al.*, 2016). The known stay-green loci, (*Stg1–4*) underlie post-flowering drought tolerance in near-isogenic lines (Harris *et al.*, 2007; Borrell *et al.*, 2014). They have been shown to confer grain yield increase, modify canopy architecture, water supply, and phenology under post-flowering drought tolerance. The alleles at *Stg1–4* are dominant and explained up to 30% of the

phenotypic variance. In this dissertation, stay-green alleles were identified in the West African sorghum and contributed up to 25% of phenotypic variance, suggesting oligogenic architecture of stay-green in the West Africa germplasm.

Several natural variants in positive pleiotropic drought response QTLs identified in chapter 4 are selected for drought tolerance improvement of elite and locally preferred cultivars (Table 5-1). These cultivars are key lines breeding programs due to their agronomic characteristics (e.g., IRAT 204, Faourou, Mota Maradi, MDK, Sorvato-1, SEPON82, SEGUIFA, Grinkan, IRAT 4, 53-49). Locally preferred cultivars used as trait donors or for improvement were selected based on their agronomic importance, popularity to growers, and favorable allele presence. These cultivars include, for example, Congossane (formerly known as IRAT 4), Sevil Ndanery (63-23), Tigne (53-37), SL 179 (50-17), Fellah, 50-16. The natural variants for drought response will be converted to Kompetitive Allele Specific PCR (KASP) markers to validate their contribution to drought tolerance in West African breeding populations. Once validated in managed water stress environments on station and in farmer fields, these markers can be routinely used as diagnostic tools to follow the introgression of drought tolerance alleles into elite backgrounds. I hypothesize that *Stg* alleles loci confers drought tolerance in elite lines of West Africa. KASP marker alleles from the drought response variants at the Stg loci would segregate in near-isogenic lines (NILs) relative to recurrent elite parents. The MABC will be conducted to introgress drought response alleles into elite and locally preferred varieties. I expect to obtain new drought-tolerant versions of locally preferred cultivars.

To take advantage of discoveries from the international community, KASP markers at (*Stg3a* and *Stg3b*) developed by the ICRISAT breeding program are being tested across West African breeding populations. The objective is to verify whether the developed KASP makers at *Stg3a* and *Stg3b* would differentiate the recurrent parental lines and B35 line for use in marker-assisted selection. There are some promising markers that segregate between B35 (donor line) and Tx7000 (susceptible line) and Senegal sorghum elite varieties. These markers will be used to monitor the introgression of stay-green alleles into elite backgrounds, Nganda, Congossane, Farourou, Darou, Nguinthe, and Kapelga.

## Marker-Assisted Backcrossing for Photoperiodic Flowering and Semi-Loose Panicle Improvement

The Senegal sorghum breeding program released high yielding new varieties that were developed to adapt to the Soudano-Sahelian zone of Senegal (semi-arid zone). These varieties are photoperiod insensitive and mature early. Growers in humid and sub-humid regions are interested in growing the new varieties. However, these varieties are not adapted to the southern region where the growing season is much longer. The varieties mature in the middle of the growing season before the end of rains leading to grain mold infection. In addition, the panicle morphology is semi-compact, which is a source of grain mold infection. The goal is to develop moderately photoperiod sensitive (medium maturity cycle) versions of these varieties with semi-loose and long panicles to adapt to the growing season length and escape grain mold infection. In the energy biomass sorghum, the dominant allele at *SbGhd7*/*Ma6* (*Maturity6*) and *Ma1* flowering time genes increases photoperiod sensitivity and delays flowering in an additive manner (Murphy *et al.*, 2014). In the sorghum conversion program, the exotic photoperiod-sensitive genotypes contained *Ma1ma6*, *ma1Ma6*, or *Ma1Ma6* (Klein *et al.*, 2008; Murphy *et al.*, 2014). Therefore, *Ma6* allele may confer moderate photoperiod sensitivity in elite lines of Senegal. In this dissertation, quantitative trait loci (QTL) at *Ma6* were identified in chapter 2 and chapter 3. The genetic variants at *Ma6* segregate between donor lines and elite backgrounds (Table 5-2). KASP markers at the *Ma6* allele will be developed to follow the introgression of photoperiodic sensitivity into elite backgrounds.

Using available whole genome resequencing data (Bellis *et al.*, 2020), I will test the hypothesis that *Ma6* allele alone confers moderate photoperiod sensitivity in elite backgrounds. The competitive hypothesis is that photoperiod sensitivity is associated with *Ma6* allele and other photoperiodic flowering genes (e.g., *Ma1–Ma5*) in elite backgrounds. To evaluate the prediction that genetic variations at *Ma6* will be strongly associated with photoperiod sensitivity in elite backgrounds across geographic regions, field trials could be performed on research stations of three different agro-ecological regions, Bambey (center-north), Nioro (center-south), and Sinthiou Maleme (south) across Senegal.

The *SP1* (*short panicle 1*) gene encodes a conserved peptide transporter 2 domain and expresses in the phloem of panicle branches to alter the panicle length and primary branch length at the basal part of the panicle (Li *et al.*, 2009). The mutant allele of *SP1* confers a short panicle

with reduced branch length. The longer the branches the looser the panicle such that loose panicle is associated with longer primary branches at basal parts of the panicle. The two natural variants in *SP1* that were identified in chapter 2 of this dissertation would be converted to KASP markers to detect genotypes with long and loose/semi-loose panicles (guinea and caudatum sorghums). The allele for longer and looser panicles are harbored by guinea and some caudatum sorghum accessions (loose and semi-loose panicles) relative to durra (compact panicle) sorghum (Table 5-3). These lines would be used for gene pyramiding to integrate *Ma6* and *SP1* alleles into elite backgrounds, including medium plant height alleles at *qHT7.1* and *Dw3* from chapter 3 (Table 5-4) and have been shown to be in repulsion linkage for height variation (Li *et al*., 2015).

## Marker-Assisted Backcrossing for Tannin Improvement

Several sorghum varieties were released to farmers by the Senegal breeding program. However, some of these varieties are characterized by high tannin content in the pericarp. Grain sorghum is characterized by a range of tannins (Dykes *et al.*, 2013), which are available as condensed tannins (proanthocyanidins) and anthocyanins (Xiong *et al.*, 2019). The tannin content in the pigmented testa is controlled by the presence of loss-of-function alleles of both *Tannin1* (*Tan1*) and *Tannin2* (*Tan2*) genes (Wu *et al.*, 2012, 2019). Functional alleles at both genes must be present to confer non-tannin. KASP markers for *Tan1* and *Tan2* alleles have been developed at Kansas State University. My research goal will be to follow the introgression of these alleles into elite backgrounds to develop tannin free versions of high tannin content varieties. The variety, CE 180-33 has a good productivity and tolerance to water-limited conditions; however, its grain quality is low with notable high tannin content, which limits its expansion in farmers' fields. The variety Nganda has a good grain quality and is highly grown by farmers but it may have some tannin trace. To develop a tannin-free version of these varieties, I hypothesize that both functional *Tan1* and *Tan2* alleles together confer non-tannin in the elite backgrounds. To evaluate this hypothesis, I predict that the KASP markers at these alleles would perfectly predict the presence of both *Tan1* and *Tan2* alleles in NILs relative to recurrent elite backgrounds.

## Genomic Selection for Grain Yield Improvement under Drought

My second research goal is to implement genomic selection, with the integration of climate- and drought-yield-associated SNPs in chapters 2 and 4 to improve genetic gain under various drought scenarios. Genomic selection (GS) would contribute to increasing genetic gain

by reducing the number of breeding cycles. In classical breeding, phenotyping is performed during the whole process of varietal development and selection. However, in GS, phenotyping is performed on the training set used to train GS models. Selection is then made based on the individual's genomic-estimated breeding values (GEBV). Thus the whole performance of individuals is obtained regardless of the genetic architecture of the traits. Moreover, GS does not require knowing the causative variants because the breeding value of individuals is estimated as the sum of genome-wide markers used via the kinship. To achieve this research goal, I hypothesize that genomic selection can efficiently increase genetic gain in three to four cycles of selection over three years while maintaining genetic diversity and limiting phenotyping efforts.

A backcross nested association mapping population (BC-NAM) is being developed from different genetic backgrounds in West Africa and elsewhere. The parents of the BC-NAM include diverse lines with important agronomic and economic values, high yielding lines, drought tolerant lines, early maturity lines, grain quality, high protein digestibility, grain mold resistance, and midge resistance lines. The Senegal NAM parental lines will also be used because the common parent, Nganda, is part of the BC-NAM. All parental lines will be intercrossed twice and self-pollinated one or two times to produce the Cycle 0 (C0) families, allowing more recombination and variability. The climate- and drought tolerance-associated alleles identified in chapter 2 and chapter 4 will be integrated into genomic selection models to increase prediction accuracy. Phenotypic measurement for yield-related traits will be carried out on the C0 families. C0 families will be used as training sets to train GS models. The best GS model will be implemented in the breeding pipeline to continuously select the best performing lines.

## References


Bellis ES, Kelly EA, Lorts CM, Gao H, DeLeo VL, Rouhan G, Budden A, Bhaskara GB, Hu Z, Muscarella R, *et al.*2020. Genomics of sorghum local adaptation to a parasitic plant. *Proceedings of the National Academy of Sciences*.

Borrell AK, van Oosterom EJ, Mullet JE, George-Jaeggli B, Jordan DR, Klein PE, Hammer GL. 2014. Stay-green alleles individually enhance grain yield in sorghum under drought by modifying canopy development and water uptake patterns. *New Phytologist* 203: 817–830.

Harris K, Subudhi PK, Borrell A, Jordan D, Rosenow D, Nguyen H, Klein P, Klein R, Mullet J. 2007. Sorghum stay-green QTL individually reduce post-flowering drought-induced leaf senescence. *Journal of Experimental Botany* 58: 327–338.

Haussmann B, Mahalakshmi V, Reddy B, Seetharama N, Hash C, Geiger H. 2002. QTL mapping of stay-green in two sorghum recombinant inbred populations. *Theoretical and Applied Genetics* 106: 133–142.

Hayes CM, Weers BD, Thakran M, Burow G, Xin Z, Emendack Y, Burke JJ, Rooney WL, Mullet JE. 2016. Discovery of a Dhurrin QTL in Sorghum: Co-localization of Dhurrin Biosynthesis and a Novel Stay-green QTL. *Crop Science* 56: 104–112.

Kebede H, Subudhi PK, Rosenow DT, Nguyen HT. 2001. Quantitative trait loci influencing drought tolerance in grain sorghum (Sorghum bicolor L. Moench). *Theoretical and Applied Genetics* 103: 266–276.

Klein RR, Mullet JE, Jordan DR, Miller FR, Rooney WL, Menz MA, Franks CD, Klein PE. 2008. The Effect of Tropical Sorghum Conversion and Inbred Development on Genome Diversity as Revealed by High-Resolution Genotyping. *Crop Science* 48: S-12-S-26.

Li S, Qian Q, Fu Z, Zeng D, Meng X, Kyozuka J, Maekawa M, Zhu X, Zhang J, Li J, *et al.*2009. Short panicle1 encodes a putative PTR family transporter and determines rice panicle size. *The Plant Journal* 58: 592–605.

Li, X., X. Li, E. Fridman, T.T. Tesso, and J. Yu. 2015. Dissecting repulsion linkage in the dwarfing gene Dw3 region for sorghum plant height provides insights into heterosis. PNAS 112:11823–11828. doi:10.1073/pnas.1509229112.

Murphy, R. L., Morishige, D. T., Brady, J. A., Rooney, W. L., Yang, S., Klein, P. E., & Mullet, J. E. (2014). Ghd7 (Ma 6) Represses Sorghum Flowering in Long Days: Ghd7 Alleles Enhance Biomass Accumulation and Grain Production. *The Plant Genome*, *7*(2). https://doi.org/10.3835/plantgenome2013.11.0040

Tuinstra MR, Grote EM, Goldsbrough PB, Ejeta G. 1996. Identification of quantitative trait loci associated with pre-flowering drought tolerance in sorghum. *Crop Science* 36: 1337–

1344.

Tuinstra MR, Grote EM, Goldsbrough PB, Ejeta G. 1997. Genetic analysis of post-flowering drought tolerance and components of grain development in Sorghum bicolor (L.) Moench. *Molecular Breeding* 3: 439–448.

Varshney RK, Mohan SM, Gaur PM, Gangarao NVPR, Pandey MK, Bohra A, Sawargaonkar SL, Chitikineni A, Kimurto PK, Janila P, *et al.*2013. Achievements and prospects of genomics-assisted breeding in three legume crops of the semi-arid tropics. *Biotechnology Advances* 31: 1120–1134.

Wu Y, Guo T, Mu Q, Wang J, Li X, Wu Y, Tian B, Wang ML, Bai G, Perumal R, *et al.*2019. Allelochemicals targeted to balance competing selections in African agroecosystems. *Nature Plants* 5: 1229–1236.

Wu Y, Li X, Xiang W, Zhu C, Lin Z, Wu Y, Li J, Pandravada S, Ridder DD, Bai G, *et al.*2012. Presence of tannins in sorghum grains is conditioned by different natural alleles of Tannin1. *Proceedings of the National Academy of Sciences of the United States of America* 109: 10281–10286.

Xu W, Rosenow DT, Nguyen HT. 2000. Stay green trait in grain sorghum: relationship between visual rating and leaf chlorophyll concentration. *Plant Breeding* 119: 365–367

**Table 5-1 SNP markers with positive pleiotropic effects across various drought scenarios and donor lines for drought tolerance improvement of elite cultivars.**

| **Locus** | ***Stg3a*** | | | ***Stg3b*** | | | | | ***Stg2*** | | | ***Stg1*** | | ***Stg4*** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SNP[a] | S2_56682379 | S2_59129283 | S2_60191986 | S2_62095163 | S2_62973945 | S2_63381610 | S2_63881780 | S2_65658140 | S3_56094063 | S3_57614567 | S3_57615696 | S3_65430305 | S3_66366589 | S5_13190947 | S5_20251208 |
| Nucleotide | C > A | A > G | A > G | C > T | C > G | C > T | T > A | C > T | C > T | A > G | T > C | C > T | T > A | G > A | G > A |
| Effect[b] | 11 | 9 | 4 | 28 | 26 | 30 | 15 | 25 | 28 | 15 | 13 | -7 | 15 | 17 | 28 |
| Lines | | | | | | | | | | | | | | | |
| IRAT 204 | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
| Faourou | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
| 53-49 | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
| Sorvato-1 | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
| Grinkan | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
| SEPON82 | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
| SRN39 | – | + | – | + | + | + | + | – | + | + | + | – | – | – | – |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MDK | – | – | – | + | + | + | + | + | – | + | – | – | – | – | – |
| Seguifa | – | + | – | + | + | + | + | + | + | + | + | – | – | + | + |
| 199SSM973D | – | – | –/+ | – | – | – | – | – | – | – | – | – | + | + | – |
| 63-23 | + | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
| IRAT 4 | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
| Tigne | – | – | – | – | – | – | – | – | – | – | – | + | – | – | – |
| SL 179 | – | – | + | – | – | – | – | – | – | – | – | – | – | – | – |
| Fellah | – | – | – | –/+ | –/+ | + | –/+ | –/+ | – | + | + | – | – | – | – |
| 50-16 | – | – | – | + | + | + | + | + | – | – | – | – | – | – | – |

[a] Single Nucleotide polymorphism where digits before and after underscore indicate chromosome number and nucleotide position on the genome, respectively; [b] effect of the alternative allele of each SNP on grain weight per panicle across diverse water stress environments

**Table 5-2. SNP markers and donor lines for early or late flowering time improvement of elite cultivars.**

| Locus | | *Ma6* | | | *SbCN8* | |
|---|---|---|---|---|---|---|
| SNP [a] | | S6_651847 | S6_697299 | S6_699843 | S9_54917833 | S9_54968379 |
| Nucleotide | | G > C | C > G | C > T | C > T | T > A |
| Effect [b] | | 30 | 29 | -7 | 29 | 29 |
| Line | DFLo | | | | | |
| Tx623 [c] | 71 [d] | – | – | – | – | – |
| Tx7000 [c] | 58 [d] | – | – | – | – | – |
| IRAT 204 | 62 | – | – | – | – | – |
| Faourou | 66 | – | – | – | – | – |
| Sorvato-1 | – | – | – | –/+ | – | – |
| MDK | 63 | – | – | – | – | – |
| Mota Maradi | 62 | – | – | – | – | – |
| 53-49 | 62 | – | – | + | – | – |
| IRAT 4 | 72 | – | – | + | – | – |
| 183SSM205G | 106 | + | + | – | + | + |
| TGVL 1 | 109 | + | + | – | + | + |
| SMIL-155 | 89 | – | – | + | – | – |

[a] Single Nucleotide Polymorphism (SNP) where digits before and after underscore indicate chromosome number and nucleotide position on the genome, respectively; [b] effect of the alternative allele of each SNP; [c] Tx623 and Tx7000 lines that have the recessive allele of *Ma6* based on Murphy *et al*. (2014); [d] days to flowering for Tx623 and Tx7000 under summer conditions based on Murphy *et al*. (2014); DFLo, days to flowering under long days of the summer in Bambey, Senegal.

**Table 5-3. SNP markers and lines segregating for marker alleles in *SP1* for panicle compactness improvement of elite cultivars.**

| | Locus | *SP1* | |
|---|---|---|---|
| | SNP[a] | S1_55302939 | S1_55305415 |
| | Nucleotide | C > T | T > C |
| Line | Type[b] | | |
| IRAT 204 | caudatum | – | – |
| Faourou | caudatum | – | – |
| PI 514461 | caudatum | – | – |
| 53-49 | guinea | – | – |
| IRAT 4 | guinea | – | – |
| PI 514342 | guinea | – | – |
| PI 514278 | durra | + | + |
| PI 514327 | durra | + | + |
| PI 514367 | durra | + | + |

[a] Digits before and after underscore indicates chromosome number and nucleotide position on the genome, respectively; [b] the durra type is associated with a short panicle and short panicle branch length relative to guinea and caudatum types.

**Table 5-4. SNP markers and lines segregating for marker alleles at *qHT7.1* and *Dw3* for plant height variation.**

| | Locus | *qHT7.1* | | *Dw3* | |
|---|---|---|---|---|---|
| | SNP | S7_56232413 | S7_56432423 | S7_59402662 | S7_59955806 |
| | Nucleotide | C > G | A > G | T > C | A > G |
| Line | PH (cm) | | | | |
| BTx623[a] | | – | – | – | – |
| Tx430[a] | | ./. | – | + | ./. |
| Tx7000[a] | | – | – | – | – |
| BTx642[a] | | – | – | – | – |
| P898012[a] | | + | + | + | + |
| Hegari[a] | | + | + | + | + |
| IRAT 204 | 132 | + | + | – | – |
| Faourou | 138 | + | + | – | – |
| Sorvato-1 | – | – | – | + | + |
| TGVL 1 | 326 | – | – | + | + |
| Mota Maradi | 197 | – | – | + | + |
| MDK | 229 | – | – | + | + |
| SEPON 82 | 130 | + | + | – | – |
| SRN39 | 148 | + | + | – | – |
| 53-49 | 270 | – | – | + | + |

The plus sign indicates that the genotype has the alternative (non-reference) nucleotide, while negative sign indicates that genotype has the reference nucleotide of BTx623; Wildtype allele, Dwarf allele;

[a] Genotype of the line based on re-sequencing data in Phytozome; PH, plant height.