Quantitative genomic analysis of agroclimatic traits in sorghum

By

Olatoye Olalere Marcus

B.Sc., Olabisi Onabanjo University, Nigeria, 2008

M.Sc., Universität Hohenheim, Germany, 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

2017

Abstract

Climate change has been anticipated to affect agriculture, with most profound effects in regions where low input agriculture is being practiced. Understanding of how plants evolved in adaptation to diverse climatic conditions in the presence of local stressors (biotic and abiotic) can be beneficial for improved crop adaptation and yield to ensure food security. Great genetic diversity exists for agroclimatic adaptation in sorghum (*Sorghum bicolor* L. Moench) but much of it has not been characterized. Thus, limiting its utilization in crop improvement. The application of next-generation sequencing has opened the plant genome for analysis to identify patterns of genome-wide nucleotide variations underlying agroclimatic adaptation.

To understand the genetic basis of adaptive traits in sorghum, the genetic architecture of sorghum inflorescence traits was characterized in the first study. Phenotypic data were obtained from multi-environment experiments and used to perform joint linkage and genome-wide association mapping. Mapping results identified previously mapped and novel genetic loci underlying inflorescence morphology in sorghum. Inflorescence traits were found to be under the control of a few large and many moderate and minor effect loci. To demonstrate how our understanding of the genetic basis of adaptive traits can facilitate genomic enabled breeding, genomic prediction analysis was performed with results showing high prediction accuracies for inflorescence traits.

In the second study, the sorghum-nested association mapping (NAM) population was used to characterize the genetic architecture of leaf erectness, leaf width, and stem diameter. About 2200 recombinant inbred lines were phenotyped in multiple environments. The obtained phenotypic data was used to perform joint linkage mapping using ~93,000 markers. The proportion of phenotypic variation explained by QTL and their allele frequencies were estimated. Common and moderate effects QTL were found to underlie marker-trait associations. Furthermore, identified QTL co-localized with genes involved in both vegetative and inflorescence development. Our results provide insights into the genetic basis of leaf erectness and stem diameter in sorghum. The

identified QTL will also facilitate the development of genomic-enable breeding tools for crop improvement and molecular characterization of the underlying genes

Finally, in a third study, 607 Nigerian accessions were genotyped and the resulting genomic data [about 190,000 single nucleotide polymorphisms (SNPs)] was used for downstream analysis. Genome-wide scans of selection and genome-wide association studies (GWAS) were performed and alongside estimates of levels of genetic differentiation and genetic diversity. Results showed that phenotypic variation in the diverse germplasm had been shaped by local adaptation across climatic gradient and can provide plant genetic resources for crop improvement.

Quantitative genomic analysis of agroclimatic traits in sorghum

By

Olatoye Olalere Marcus

B.Sc., Olabisi Onabanjo University, Nigeria, 2008

M.Sc., Universität Hohenheim, Germany, 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

2017

Approved by:

Major Professor
Geoffrey P. Morris

Copyright

Abstract

Climate change has been anticipated to affect agriculture, with most the profound effect in regions where low input agriculture is being practiced. Understanding of how plants evolved in adaptation to diverse climatic conditions in the presence of local stressors (biotic and abiotic) can be beneficial for improved crop adaptation and yield to ensure food security. Great genetic diversity exists for agroclimatic adaptation in sorghum (*Sorghum bicolor* L. Moench) but much of it has not been characterized. Thus, limiting its utilization in crop improvement. The application of next-generation sequencing has opened the plant genome for analysis to identify patterns of genome-wide nucleotide variations underlying agroclimatic adaptation.

To understand the genetic basis of adaptive traits in sorghum, the genetic architecture of sorghum inflorescence traits was characterized in the first study. Phenotypic data were obtained from multi-environment experiments and used to perform joint linkage and genome-wide association mapping. Mapping results identified previously mapped and novel genetic loci underlying inflorescence morphology in sorghum. Inflorescence traits were found to be under the control of a few large and many moderate and minor effect loci. To demonstrate how our understanding of the genetic basis of adaptive traits can facilitate genomic enabled breeding, genomic prediction analysis was performed with results showing high prediction accuracies for inflorescence traits.

In the second study, the sorghum-nested association mapping (NAM) population was used to characterize the genetic architecture of leaf erectness, leaf width, and stem diameter. About 2200 recombinant inbred lines were phenotyped in multiple environments. The obtained phenotypic data was used to perform joint linkage mapping using ~93,000 markers. The proportion of phenotypic variation explained by QTL and their allele frequencies were estimated. Common and moderate effects QTL were found to underlie marker-trait associations. Furthermore, identified QTL co-localized with genes involved in both vegetative and inflorescence development. Our results provide insights into the genetic basis of leaf erectness and stem diameter in sorghum. The identified QTL will also facilitate the development of genomic-enable breeding tools for crop improvement and molecular characterization of the underlying genes

Finally, in a third study, 607 Nigerian accessions were genotyped and the resulting genomic data [about 190,000 single nucleotide polymorphisms (SNPs)] was used for downstream analysis. Genome-wide scans of selection and genome-wide association studies (GWAS) were performed and alongside estimates of levels of genetic differentiation and genetic diversity. Results showed that phenotypic variation in the diverse germplasm had been shaped by local adaptation across climatic gradient and can provide plant genetic resources for crop improvement.

Table of Contents

List of Figures

List of Tables

Acknowledgement

I am thankful to God for blessing me with the privilege to work with Dr. Geoffrey Morris, a man who wasn’t just a supervisor but a great mentor as well. I will like to thank Dr. Ramasamy Perumal for his tremendous efforts in always providing me with all the resources needed to conduct my fieldwork at the Kansas State University Agricultural Research Center, Hays, Kansas. I will also like to thank my committee members Dr. Kathrin Schrick, Dr. Mary Beth Kirkham and Dr. Jesse Poland for invaluable advice and guidance.

My sincere appreciation goes to all my lab members Sophie Bouchet, Sandeep Marla, Zhenbin Hu, Jianan Wang, Kebede Muleta, Sridevi Nakka, Terry Felderhoff, Brian Wempen, Kelly Qinling Li, Fanna Maina, Jacques Faye, Matt Davis, Rich Brown, Ruth Bartel, and Obembe Oladipo.

To my wife and helpmeet, Deborah Isimemen Olatoye, thank you for your support all along. And to my son Eleazer Olatoye sorry for the times I had to leave you at home even when you wanted to play with me. I love you guys.

My appreciation also goes to the Westview Wesleyan community, the Wolters’, Cranfords, Nelsons, Hendricks, McHenrys and Smiths.

Dedication

To the Lord God and giver of life, to the memory of innocent students and National Youth Service Corps members massacred by Islamic extremists in Northern Nigeria during the 2011 post election violence, and to the unknown godly Moslem man that gave me a ride in the midst of the chaos. I hold no prejudice.

Chapter 1 – Agroclimatic adaptation in crops species

Climatic adaptation

Organisms have adapted to a wide range of ecological and geographic environments. Adaptation of a particular specie is dependent on the movement of its population towards a optimal phenotype, which best suits that specific environment (Orr, 2005). Adaptation to diverse environments results in intraspecific variation of morphological and physiological traits. The major climatic factors influencing trait adaptation in biological systems are ultraviolet (UV) radiation, photoperiodicity, temperature, and precipitation. These factors shape the organismal response to UV intensity, disease, heat, cold and drought. Climatic adaptation has been reported in both model and non-model systems (Table 1.1).

Agroclimatic adaptation in crop species

Climate adaptation has been observed in crop species (Harlan, 1992). For instance, latitudinal flowering time adaptation in maize (Camus-Kulandaivelu *et al.*, 2006; Buckler *et al.*, 2009; Romero Navarro *et al.*, 2017) and inflorescence variation due to agroclimatic adaptation in sorghum (de Wet & Huckabay, 1967; Harlan, 1992) are classic examples. In crop species, extensive genetic diversity exists for adaptive phenotypes that have been subjected to not only a natural, but also artificial (human) selection in response to local climatic conditions across diverse environments (Lasky *et al.*, 2015). Landraces (traditional varieties) (Zeven, 1998) are typically adapted to local stress (biotic and abiotic) and are valuable plant genetic resources for crop improvement (Lasky *et al.*, 2015). The resulting phenotypic divergence of traits in response to local adaptation across climatic zones is regarded as agroclimatic trait variation (Figure 1.1). Agroclimatic adaptive traits are often confounded with population structure (Camus-Kulandaivelu *et al.*, 2006; Valdar *et al.*, 2009; Samis *et al.*, 2012; Morris *et al.*, 2013; Bouchet *et al.*, 2017), therefore it can be difficult to characterize their genetic architecture.

Genetic architecture of agroclimatic traits

A neo-Darwinian perspective of adaptation was given by Fisher stating that an allelic variant either from novel mutation or standing variation will increase in frequency and become fixed as an adaptive response to environmental changes (Hughes, 2012). Adaptive evolution is mostly dependent on polygenic characters (Lande & Barrowclough, 1996). The Fisher-Orr model of adaptation (Orr, 1998, 2005) describes the changing genetic architecture of complex adaptive traits using a geometric model (Orr, 1999, 2005). The model describes a pattern of diminishing returns, as few large effect loci are fixed first, followed by numerous small effect loci. One consequence of the model is that for a trait that is close to its fitness optimum, only small effect loci can bring it closer to the optimum (Orr, 2005). The genetic architecture entails the number of quantitative trait loci (QTL), allele frequency, the effect size of the QTL, identity of the genes and gene actions associated with a particular trait (Mackay, 2001). It is important for response to selection and can provide insight about its evolutionary history (Brown *et al.*, 2011). Large effect loci have been found to underlie adaptive traits in *Sorghum bicolor* (Lin *et al.*, 1995; Bouchet *et al.*, 2017), while small effect loci were found to underlie adaptive traits in maize (*Zea mays*) (Buckler *et al.*, 2009; Peiffer *et al.*, 2014). For instance, small effect size loci (with an effect size of < 1 day) associated with flowering time have been identified in maize and sorghum (Buckler *et al.*, 2009; Bouchet *et al.*, 2017).

Furthermore, the expression of any complex trait resulting from developmental or chemical pathways is often comprised of a network of loci interacting at both genetic and molecular levels (Mackay, 2001). These genotype-by-genotype interactions are known as epistasis. In classical quantitative genetics, epistasis describes non-additivity of effects at multiple loci and exhibits various levels of complexity of interactions depending on the numbers of loci being considered (Lynch & Walsh, 1998). In a two loci model, there will be three levels of interactions: additive-by-additive, additive by dominance and dominance-by-dominance forms of interactions. However, in experiments using recombinant inbred lines (RILs) only the additive-by-additive interaction will be observed since there are only two homozygous classes of alleles. Epistatic interactions of loci have been associated with flowering time in Arabidopsis (Juenger *et al.*, 2005) and maize (Buckler *et al.*, 2009).

Quantitative genomic dissection of agroclimatic adaptation

Quantitative trait dissection is the mapping of genetic regions responsible for a trait through marker-phenotype association (Lander & Schork, 1994). Most of the QTL identified in trait mapping in crops has been based on bi-parental mapping populations, either F2 individuals or recombinant inbred lines (RILs) generated from the cross between two parents contrasting for the trait of interest (Mackay *et al.*, 2009). However, only QTL associated with the phenotypic variation generated from the controlled cross between the two parents can be identified. Thus, only a small fraction of the genetic diversity available for a particular species are captured (Myles *et al.*, 2009). Additionally, due to small population sizes (< 250), the effect sizes of identified QTL are usually inflated (known as the Beavis effect) (Utz *et al.*, 2000; Juenger *et al.*, 2005). Also due to limited recombination, the QTL regions are usually large (Myles *et al.*, 2009). Conversely, genome-wide association mapping, also known as linkage disequilibrium (LD) mapping exploits, LD between genotyped markers and functional variants/polymorphisms associated with phenotypic differences caused by the polymorphism (Ehrenreich & Purugganan, 2006). Association mapping populations (or diversity panels) offers increased mapping resolution and diversity for mapping due to historical recombination. The level of resolution with which a trait can be mapped depends on the extent of LD in a given genome region (Myles *et al.*, 2009). However, population structure in the population leads to spurious associations, thereby limiting mapping power.

Population structure reflects genome-wide non-random association of alleles and correlation between allele frequencies and phenotypic variation (Myles *et al.*, 2009). Population structure may lead to correlations between phenotypic variation and genetic relatedness. This results in spurious genotype-phenotype covariance, which makes genome-wide markers appear to be associated with the trait. This has been demonstrated by fitting a naive model (which does not account for population structure) to perform genome-wide association mapping of adaptive traits (Huang *et al.*, 2010; Morris *et al.*, 2013). Models accounting for population structure have been established and resulted in the identification of true associations (Yu *et al.*, 2006; Zhang *et al.*, 2010; Korte *et al.*, 2012; Segura *et al.*, 2012). However, for adaptive traits, the model accounting for structure can be biased against true variants associated with population structure, leading to false negatives (Myles *et al.* 2009; Lipka *et al.* 2015). This bias limits the characterization of genetic architecture underlying agroclimatic traits.

Multi-parental populations were designed to overcome the challenge associated with historical population structure. Examples include Multi-Advanced Generation Intercross (MAGIC) and Nested Association Mapping populations (NAM). The NAM design involves a cross between a common founder and a set of diverse founders, followed by selfing of the F_1 for several generations to generate RILs (Paterson, 2013). NAM populations have been used to characterize the genetic architecture of adaptive traits in maize and sorghum (Buckler *et al.*, 2009; Brown *et al.*, 2011; Peiffer *et al.*, 2014; Bouchet *et al.*, 2017). In NAM populations, increased LD within families facilitates haplotype imputation for missing data when using low coverage genomic data (Brachi *et al.*, 2011).

Population genomic analysis of agroclimatic adaptation

The availability of next-generation sequencing technologies has made possible genome-wide analysis of patterns of selection using population genomics in plants. The effects of selection on the pattern of genome-wide nucleotide variation vary. These are the extent of linkage disequilibrium around regions under selection, amount and structure of polymorphism, the degree of population differentiation, and the rate and percentage of nonsynonymous substitution (Siol *et al.*, 2010). The methods used to explore patterns of nucleotide variation have been categorized into two categories: first, those that identify fingerprints of selection on linked neutral loci and second, those that deduce the action of selection on the loci (Siol *et al.*, 2010). These approaches have been applied to genomic data of plants sampled across diverse climatic regions, both for biological model species and agricultural crop species. Genome scans have been used to identify molecular targets of selection (Olsen *et al.*, 2006; Gore *et al.*, 2009; Ishii *et al.*, 2013; Yoder *et al.*, 2014; Gouesnard *et al.*, 2017). In addition, climatic variables have been used as proxies for unknown traits underlying agroclimatic adaptation in GWAS (Hancock *et al.*, 2011; Fournier-Level *et al.*, 2011; Horton *et al.*, 2012; Yoder *et al.*, 2014; Lasky *et al.*, 2015). Some studies have integrated quantitative genomic approaches with population genomic approaches to dissect the genetic basis of climatic adaptation (Morris *et al.*, 2013; Yoder *et al.*, 2014; Lasky *et al.*, 2015; Gouesnard *et al.*, 2017).

Conclusion

Quantitative trait genomics, population genomics, and integrative genomic approaches can improve our understanding of crop adaptation and facilitate conservation and utilization of crop germplasm (Morris *et al.*, 2013; Hu *et al.*, 2015; Lasky *et al.*, 2015). This will result in an increase in the amount of genetic diversity utilized in crop improvement (Morris *et al.*, 2013), facilitate the characterization of the genetic architecture of adaptive traits and identification of underlying genes for genomic-enabled breeding applications. These will be helpful in low input agricultural systems of the world where the impact of climate change is expected to be greatest.

Tables and Figures

Table 1.1: Selected list of traits responsible for clinal adaptation

Species	Adaptation	Literature
Arabidopsis thaliana	Flowering time	Samis et al. 2012
Arabidopsis thaliana	Seed dormancy	Kronholm et al. 2012
Arabidopsis thaliana	Freezing tolerance	Zhen et al. 2008
Drosophila melanogaster	Ultraviolet intensity	Bastide *et al.* 2016
Mus domesticus	Cold tolerance	Lynch 1992
Engraulis encrasicolus	Thermal adaptation	Harlan 1992

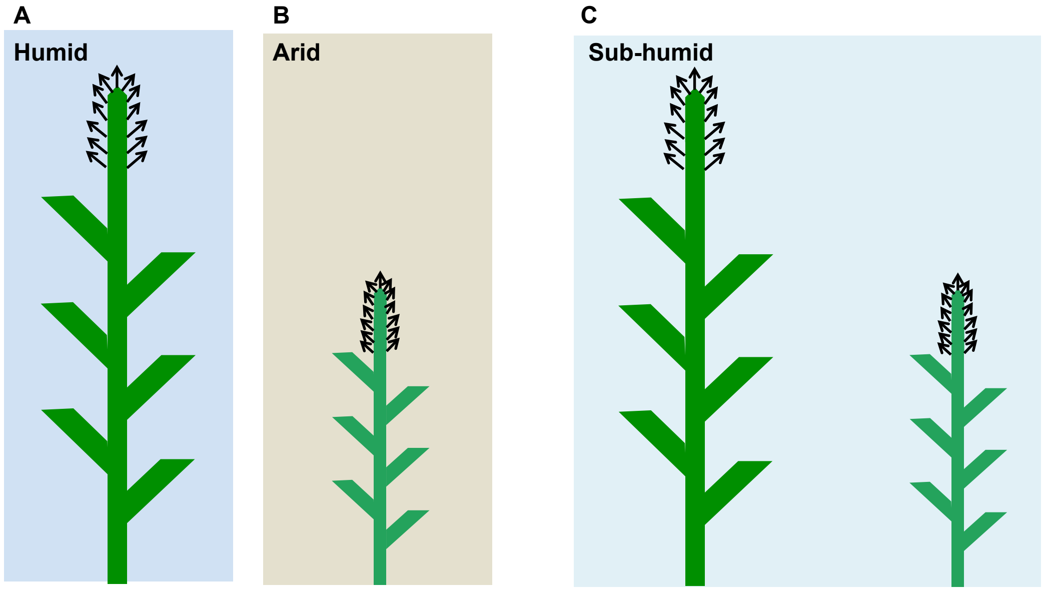

Figure 1.1: Climatic variation in plant architecture.

Plants of the same species in (A) humid climate and (B) arid climate with different architecture due to adaptation to different climatic conditions over a long period of time. Both plants maintain their natural architecture when planted in a common garden (C) sub-humid climate.

References

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.

Brachi B, Morris GP, Borevitz JO. **2011**. Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biology* **12**: 232.

Brown PJ, Upadyayula N, Mahone GS, Tian F, Bradbury PJ, Myles S, Holland JB, Flint-Garcia S, McMullen MD, Buckler ES, *et al.* 2011. Distinct genetic architectures for male and female inflorescence traits of maize. *PLoS Genetics* **7**: e1002383.

Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, Ersoz E, Flint-Garcia S, Garcia A, Glaubitz JC, *et al.* 2009. The Genetic Architecture of Maize. *Science* **325**: 714–718.

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**: 2449–63.

Ehrenreich IM, Purugganan MD. **2006**. The molecular genetic basis of plant adaptation. *American Journal of Botany* **93**: 953–962.

Fournier-Level A, Korte A, Cooper MD, Nordborg M, Schmitt J, Wilczek AM. **2011**. A Map of Local Adaptation in Arabidopsis thaliana. *Science* **334**: 86–89.

Gore MA, Chia J-M, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, Peiffer JA, McMullen MD, Grills GS, Ross-Ibarra J, *et al.* 2009. A first-generation haplotype map of maize. *Science* **326**: 1115–7.

Gouesnard B, Negro S, Laffray A, Glaubitz J, Melchinger A, Revilla P, Moreno-Gonzalez J, Madur D, Combes V, Tollon-Cordet C, *et al.* 2017. Genotyping-by-sequencing highlights original diversity patterns within a European collection of 1191 maize flint lines, as compared to the maize USDA genebank. *Theoretical and Applied Genetics* **130**: 2165–2189.

Hancock AM, Brachi B, Faure N, Horton MW, Jarymowycz LB, Sperone FG, Toomajian C, Roux F, Bergelson J. **2011**. Adaptation to Climate Across the Arabidopsis thaliana Genome. *Science* **334**: 83–86.

Harlan J. **1992**. *Crops & Man* (G Peterson, S Baenziger, and R Dinauer, Eds.). Madison, WI, U.S.A: American Society of Agronomy.

Horton MW, Hancock AM, Huang YS, Toomajian C, Atwell S, Auton A, Muliyati NW, Platt A, Sperone FG, Vilhjálmsson BJ, *et al.* 2012. Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. *Nature Genetics* **44**: 212–216.

Hu Z, Mbacké B, Perumal R, Guèye MC, Sy O, Bouchet S, Prasad PVV, Morris GP. **2015**. Population genomics of pearl millet (Pennisetum glaucum (L.) R. Br.): Comparative analysis of global accessions and Senegalese landraces. *BMC Genomics* **16**:

1048.

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.

Hughes A. **2012**. Evolution of adaptive phenotypic traits without positive Darwinian selection. *Heredity* **10897**: 347–353.

Ishii T, Numaguchi K, Miura K, Yoshida K, Thanh PT, Htun TM, Yamasaki M, Komeda N, Matsumoto T, Terauchi R, *et al.* 2013. OsLG1 regulates a closed panicle trait in domesticated rice. *Nature genetics* **45**: 462–5, 465–2.

Juenger TE, Sen S, Stowe KA, Simms EL. **2005**. Epistasis and genotype-environment interaction for quantitative trait loci affecting flowering time in Arabidopsis thaliana. In: Genetics of Adaptation. Berlin/Heidelberg: Springer-Verlag, 87–105.

Korte A, Vilhjálmsson BJ, Segura V, Platt A, Long Q, Nordborg M. **2012**. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature genetics* **44**: 1066–71.

Lande R, Barrowclough FG. **1996**. *Viable Populations for Conservation* (Michael E. Soule, Ed.). New York: Cambridge University Press.

Lander ES, Schork NJ. **1994**. Genetic Dissection of Complex Traits. *Source: Science, New Series Genome Issue* **26513018**: 2037–2048.

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–e1400218.

Lin Y-R, 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, Kandianis CB, Hudson ME, Yu J, Drnevich J, Bradbury PJ, Gore MA. **2015**. From association to prediction: statistical methods for the dissection and selection of complex traits in plants. *Current Opinion in Plant Biology* **24**: 110–118.

Lynch M, Walsh B. **1998**. *Genetics and Analysis of Quantitative Traits*. Sinauer.

Mackay TFC. **2001**. The Genetic Architecture of Quantitative Traits. *Annual Review of Genetics* **35**: 303–339.

Mackay TFC, Stone EA, Ayroles JF. **2009**. The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics* **10**: 565–577.

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**: 453–8.

Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, Costich DE, Buckler ES. **2009**. Association Mapping: Critical Considerations Shift from Genotyping to Experimental Design. *The Plant Cell* **21**: 2194–2202.

Olsen KM, Caicedo AL, Polato N, Mcclung A, Mccouch S, Purugganan MD. **2006**. Selection Under Domestication: Evidence for a Sweep in the Rice Waxy Genomic Region. *Genetics* **173**: 975–983.

Orr HA. **1998**. The Population Genetics of Adaptation: The Distribution of Factors Fixed during Adaptive Evolution. *Evolution* **52**: 935–949.

Orr HA. **1999**. The evolutionary genetics of adaptation : a simulation study. *Genet. Res., Camb* **74**: 207–214.

Orr HA. **2005**. The genetic theory of adaptation: a brief history. *Nature Reviews Genetics* **6**: 119–127.

Paterson AH. **2013**. Genomics of the Saccharinae. *Genomics of the Saccharinae*: 1–567.

Peiffer JA, Romay MC, Gore MA, Flint-Garcia SA, Zhang Z, Millard MJ, Gardner CAC, McMullen MD, Holland JB, Bradbury PJ, *et al.* 2014. The genetic architecture of maize height. *Genetics* **196**: 1337–1356.

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.

Samis KE, Murren CJ, Bossdorf O, Donohue K, Fenster CB, Malmberg RL, Purugganan MD, Stinchcombe JR. **2012**. Longitudinal trends in climate drive flowering time clines in north american arabidopsis thaliana. *Ecology and Evolution* **2**: 1162–1180.

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.

Siol M, Wright SI, Barrett SCH. **2010**. The population genomics of plant adaptation. *New Phytologist* **188**: 313–332.

Utz HF, Melchinger AE, Schön CC. **2000**. Bias and sampling error of the estimated proportion of genotypic variance explained by quantitative trait loci determined from experimental data in maize using cross validation and validation with independent samples. *Genetics* **154**: 1839–1849.

Valdar W, Holmes CC, Mott R, Flint J. **2009**. Mapping in structured populations by resample model averaging. *Genetics* **182**: 1263–77.

de Wet JM, Huckabay JP. **1967**. The Origin of Sorghum Bicolor. II. Distribution and Domestication. *Evolution* **21**: 787–802.

Yoder JB, Stanton-Geddes J, Zhou P, Briskine R, Young ND, Tiffin P. **2014**. Genomic Signature of Adaptation to Climate in Medicago truncatula. *Genetics* **196**: 1263–1275.

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, *et al.* 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* **38**: 203–208.

Zeven AC. **1998**. Landraces: A review of definitions and classifications. *Euphytica* **104**: 127–139.

Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, *et al.* 2010. Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics* **42**: 355–360.

Chapter 2 – The Genetic Architecture of Inflorescence Morphology in Sorghum

Abstract

The morphology of sorghum inflorescence enables agroclimatic adaptation due to variations in its compactness across arid to sub-humid climates. In this study, the genetic basis of sorghum inflorescence traits including lower branch length (LBL), upper branch length (UBL), rachis length (RL), and rachis diameter (RD) were characterized using a nested association mapping (NAM) population. Phenotypic data were obtained from multi-environment experiments and used to perform joint linkage and genome-wide association mapping. Mapping results identified previously identified and novel loci underlying inflorescence traits in sorghum. Inflorescence traits were found to be under the control of mainly minor effect loci. Quantitative trait loci (QTL) co-localized with homologs of rice and maize inflorescence genes including *ramosa2*, *sparse inflorescence1*, and *YUCCA5*. Lack of colocalization between the lower and upper inflorescence branch QTL suggest they are under independent genetic control. Some of the associations that were regarded as false negatives in GWAS models for mapping lower branch length were identified as true associations in the joint linkage mapping result. Finally, cross-validation results showed high prediction accuracies for LBL, UBL, RL, and RD. This study provided an understanding of the genetic architecture of sorghum inflorescence and prospects for genomic-enabled breeding.

Introduction

Adaptive evolution has shaped the genetic architecture of complex traits. Adaptive traits with phenotypic divergence across climatic zones (Morris *et al.*, 2013) can be referred to as agroclimatic traits. However, agroclimatic divergence often results in genetic differentiation and population structure that hinders effective characterization of trait genetic architecture and reduce genomic prediction accuracy. The genetic architecture of a trait is defined by the number, effect size, allele frequency, and gene action of the loci controlling it (Bouchet *et al.*, 2017). The Fisher-Orr model hypothesized that only loci of small effects can bring an organism close to its fitness optimum. A situation where few large effect loci are fixed first followed by numerous small effect loci. Different effect sized loci have been identified to be associated with adaptive traits in organisms. Large effect sized variants were implicated in loss of body armor in Alaskan threespine stickleback populations (Cresko *et al.*, 2004); large effect loci underlie cuticle melanin variation in African *Drosophila melanogaster* (Bastide *et al.*, 2016); and inflorescence architecture in maize and rice are under the control of large and small effect loci (Brown *et al.*, 2011; Crowell *et al.*, 2016). Flowering time is controlled by numerous small effect loci in maize, an outcrosser, in contrast to large effect loci in selfing species like Arabidopsis, sorghum, and rice. Thus, genetic architecture is also influenced by reproductive biology (Buckler *et al.*, 2009; Brown *et al.*, 2011). The characterization of genetic architecture is dependent on the efficiency and power of trait genetic dissection.

Earlier mapping approaches utilized bi-parental populations. However, the efficiency of these populations for mapping is hindered by limited genetic diversity and small sample sizes that causes inflation of effect sizes. Genome-wide association studies (GWAS) offer a higher mapping resolution through historical recombination and higher genetic variation and resolution for mapping complex traits (Myles *et al.*, 2009). However, because agroclimatic traits are often confounded by population structure, GWAS can be limited in power to find true associations (Myles *et al.*, 2009; Huang *et al.*, 2010; Morris *et al.*, 2013). Multi-parental populations such as Multiple Advanced Generation Intercross (MAGIC) and Nested Association Mapping (NAM) (Rakshit *et al.*, 2012; Paterson, 2013; Wallace *et al.*, 2014) exhibit increased genetic diversity and mapping resolution than bi-parental populations (Bouchet *et al.*, 2017). In addition, they reduce the confounding of phenotypic variation with ancestral population structure and have more balanced allele frequencies for increased statistical power for mapping. NAM has been used to dissect quantitative traits such as for flowering time (Buckler *et al.*, 2009), inflorescence morphology (Brown *et al.*, 2011; Wu *et al.*, 2016), and height (Peiffer *et al.*, 2014)in maize and barley for flowering time (Maurer *et al.*, 2015). Recently, a sorghum NAM was developed and shown to be effective for dissection of the genetic architecture underlying agroclimatic complex traits (Paterson, 2013; Bouchet *et al.*, 2017).

Sorghum is a source of food, feed, and bioenergy in many parts of the world, especially in the semi-arid regions where maize and rice cannot thrive. In developing countries, sorghum is predominantly cultivated by smallholder farmers (National Research Council, 1996). The cultivated sorghum varieties are in most cases the locally preferred and adapted types. Sorghum has diffused to different agroclimatic zones with variations in traits such as height, leaf architecture, and inflorescence architecture through balancing selection. In particular, sorghum inflorescence varies in compactness across agroclimatic zones and play functional roles in yield components (Brown *et al.*, 2006; Witt Hmon *et al.*, 2013). In humid climates, the open inflorescence allows air movement among seeds to prevent grain damage. In the sub-Saharan African region (SSA), breeding programs are usually made up of germplasm of varied racial origins and inflorescence types. Furthermore, the application of genomic-enabled breeding methodologies in sorghum in the SSA is not well established either due to infrastructure or inadequate knowledge of the genetic basis of traits. Therefore, it is essential to characterize the genetic basis of sorghum inflorescence in order to utilize marker-assisted selection (MAS) approach for crop improvement. In addition, understanding the genetic basis of sorghum inflorescence can provide information about associated genes thus serving as a basis for further molecular characterization.

Until now, the genetic architecture of sorghum inflorescence morphology is poorly understood, and no genes have been characterized (Brown *et al.*, 2006; Witt Hmon *et al.*, 2013). However, as grass inflorescences share a close homology, studies employing mutation and other cloning approaches have identified numerous genes controlling inflorescence morphology in maize, rice and barley. Many of these genes share homology across different grass species and are involved in regulatory functions, hormone metabolism, and transport (Kellogg, 2007; Tanaka *et al.*, 2013). Some of the genes that have been characterized are maize *Sparse inflorescence1* [*Spi1*; (Gallavotti *et al.*, 2008)], rice *liguless1* [*OsLg1*; (Ishii *et al.*, 2013)], rice *sped1-D* (Jiang *et al.*, 2014), maize *branch angle defective1* (*BAD1*) (Bai *et al.*, 2012) maize *branched silkless1* (*bd1*) (Chuck *et al.*, 2002), and maize *ramosa* genes (*ra1, ra2, ra3)* (Satoh-Nagasawa *et al.*, 2006). Conserved functions of inflorescence genes have been shown among many grass species (Huang *et al.*, 2017). The objective of this study was to characterize the genetic basis of inflorescence traits in sorghum in terms of numbers of QTL, their effect size, allele frequencies and genes underlying them. Genomic dissection was performed on the sorghum NAM population derived from a cross between an elite common parent (RTx430) and 10 diverse founders (Table 2.1). These founders originated from different agroclimatic zones, thereby capturing a wide genetic and morphological diversity.

Materials and Methods

Plant Materials and Phenotyping

The sorghum NAM population is comprised of ten diverse parents (Table 2.1) and one common parent, RTx430, which is an elite breeding line. Each diverse parent and its RILs represent a family of 250 RILs making a total of 2500 RILs in the whole population. The NAM RILs were phenotyped at $F_{6:7}$ and $F_{6:8}$ generations for upper primary branch length (UBL), lower primary branch length (LBL), and rachis length (RL). All traits were phenotyped in semi-arid (Hays, Kansas) and humid continental (Manhattan, Kansas) environments for two years (2014 and 2015). A single location/site by year was regarded as one environment (Table 2.2). In the second year (2015), NAM RILs were randomized within maturity blocks of families in a row-column design based on data from the first year flowering data. Each row (corresponding to a plot) was 3 m long with 1 m alleys between ranges. Three sorghum panicles were harvested after physiological maturity per row (RIL) and subsequently used for phenotyping. Inflorescence morphology traits were measured using barcode rulers (1 millimeter precision) and barcode readers (Motorola Symbol CS3000 Series Scanner, Chicago IL, USA). Rachis length (RL) was measured on three randomly harvested panicles from a plot as the distance from apex of the panicle to the point of attachment of the lowest rachis lower primary branch (Brown *et al.*, 2006; Pann *et al.*, 2014). Rachis diameter (RD) was measured with a digital vernier caliper as the diameter of the peduncle at the point of attachment of the bottommost rachis lower primary branch. For UBL and LBL, six primary branches were randomly selected and carefully detached from the upper (UBL) and lower (LBL) regions of two panicles. The measured traits are illustrated in Figure 2.1A.

Genomic Data Analysis

Previously generated NAM population genomic data (Bouchet *et al.* 2017) were combined with a global sorghum panel to develop a large SNP data set of 14,440 total accessions. Sequence reads were aligned to the BTx623 reference genome version 3 using Burrow Wheeler Aligner 4.0 and TASSEL 5.0 (Glaubitz *et al.*, 2014) was used for SNP calling. Missing data imputation was done in two stages. The sorghum NAM population and the sorghum association mapping population (SAP) GBS data were first extracted from the build. This data was first filtered to remove triallelic SNPs, followed by filtering to remove markers missing in more than 80% of the individuals, and filtering to keep markers with > 3% minor allele frequency prior to imputation. The NAM population and SAP were each imputed separately using Beagle 4 (Browning & Browning, 2013). After imputation, the two GBS datasets (NAM and SAP) were filtered at MAF of 0.05 and RILs with more than 10 percent heterozygosity were dropped from the NAM data. The effect of SNP variants were inferred by SnpEFF program (Cingolani *et al.*, 2012) for the imputed NAM genomic data. Linkage disequilibrium (LD) decay was estimated using (BGI-shenzhen, 2017) and plotted in R.

Phenotype and Heritability Analysis

Phenotypic data analysis was carried out using R programming language and SAS (SAS Institute Inc., Cary, NC, USA). All traits were tested for normality and the only trait with skewed distribution (UBL) was log transformed. Analysis of variance was performed for each trait and the Pearson pairwise correlation between traits was also performed using R. The best linear unbiased prediction (BLUP, for data from five environments) of each trait was estimated using *lmer* function in *LME4* package in R (Bates *et al.*, 2017) with genotype, environment, and genotype-environment interactions fitted as random effects (Wu *et al.*, 2016). A linear model was fitted with traits' BLUPs and family effect and the residuals were used for pairwise correlations between traits. The variance components used for broad sense heritability estimation were analyzed using the maximum likelihood method by PROC VARCOMP of the SAS software (SAS Institute Inc., Cary, NC, USA) by fitting RILs nested within families, RIL nested within families by environments interaction as random effects (Equation 1). The resulting variance components were used to estimate the broad sense heritability following equation (3) in (Hung *et al.*, 2012). The broad sense heritability on line mean basis were estimated as:

$$H^2 = \frac{\hat{\sigma}^2_{RIL(family)_p}}{\hat{\sigma}^2_{RIL(family)_p} + \frac{\hat{\sigma}^2_{env*RIL(family)_p}}{n_{envl_p}} + \frac{\hat{\sigma}^2_{\varepsilon}}{n_{plot_p}}} \qquad [1]$$

where $\hat{\sigma}^2_{RIL(family)_p}$ is the variance component of RIL nested within family p, n_{envl_p} is the harmonic mean of the number of environments in which each RIL was observed, and n_{plot_p} the harmonic mean of the total number of plots in which each RIL. Also, Pearson pairwise correlation between traits was estimated using the residuals derived from fitting a linear model for family and trait phenotypic means;

$$\mathbf{y} = \mu + \gamma_i + \varepsilon_{ij} \qquad [2]$$

where $\mathbf{y}$ is the vector of phenotypic data, γ_i is the term for the NAM families, and ε_{ij} is the residual term.

Joint Linkage Mapping

Joint linkage analysis was performed using 92,391 markers and 2220 RILs. This approach is based on forward inclusion and backward elimination stepwise regression approaches implemented in TASSEL 5.0 (Glaubitz *et al.*, 2014) stepwise plugin and the family effect was accounted for as a co-factor in the analysis. First, a nested joint linkage (NJL) model was fitted where markers were nested within families, due to the fact that it has been found to be effective for estimating QTL effects within families (Poland *et al.*, 2011; Würschum *et al.*, 2011). In addition, a non-nested joint linkage model (JL) where markers were not nested within families was used for analysis due to its higher predictive power than NJL (Würschum *et al.*, 2011). Entry and exit F_{test} values were set to 0.001 and based on 100 permutations, the *P*-value threshold was set to 1.84 E-6. One important advantage of joint linkage mapping is that it enables effective mapping of small effect and low frequency QTL that may be missed in GWAS. The JL model was specified as;

$$y = b_o + \alpha_f u_f + \sum_{i=1}^{k} x_i b_i + e_i \qquad [3]$$

where b_0 is the intercept, u_f is the effect of the family of founder *f* obtained in the cross with the common parent (RTx430), α_f is the coefficient matrix relating u_f to *y*, b_i is the effect of the i^{th} identified locus in the model, x_i is the incidence vector that relates b_i to *y* and *k* is the number of significant QTL in the final model (Yu *et al.*, 2008).

Genome-wide Association Studies

Genome-wide association study (GWAS) was performed for all traits using 92,391 markers and 2220 RILs; first, for single environment and second, using BLUP adjusted by environments. The multi-locus-mixed model (MLMM) approach (Segura *et al.*, 2012) implemented in R was used for GWAS. The MLMM approach performs stepwise regression involving both forward and backward regressions, accounts for major loci and reduces the effect of allelic heterogeneity. The family effect was fitted as a co-factor and a random polygenic term (kinship relationship matrix) was also accounted for in the MLMM model. A total of 92,391 SNPs were used in the GWAS analysis and coded as 2 and 0 for homozygous SNPs and 1 for heterozygous SNPs. Bonferroni correction of E - 07 (α/total number of markers [5.4 E-07]; where α = 0.05) was used to determine the cut-off threshold for each trait association. Furthermore, GWAS was also performed on sorghum association diversity panel (SAP, consisting of about 334 accessions (Casa *et al.*, 2008)) using Generalized Linear Model (GLM) and Compressed Mixed Linear Model (CMLM) using GAPIT package in R (Lipka *et al.*, 2012). The GLM (naive model) did not account for population structure and was specified as;

$$\mathbf{y} = \mathbf{S}\boldsymbol{\alpha} + \mathbf{e} \qquad [4]$$

where $\mathbf{y}$ is the vector of phenotypes, $\boldsymbol{\alpha}$ is a vector of SNP effects, and $\mathbf{e}$ is the vector of residual effects. And **S** is the incident matrix of 1s and 0s relating y to $\boldsymbol{\alpha}$. The CMLM model (full model) accounted for population structure and polygenic background effects (kinship) was specified as;

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Qv} + \mathbf{Zu} + \mathbf{e} \qquad [5]$$

where $\mathbf{y}$ is the vector of phenotypic information, $\boldsymbol{\beta}$ is a vector of fixed effects other than SNP or population structure effects, $\mathbf{v}$ is a vector of fixed effects for population structure, u is an unknown vector of random additive genetic effects from multiple background QTL for RILs. **X**, **Q***,* and **Z** are incident matrices of 1s and 0s relating $\mathbf{y}$ to $\boldsymbol{\beta}$ and $\mathbf{u}$ (Yu *et al.*, 2006). The phenotypic data used for GWAS in the SAP had been previously described and published (Brown *et al.*, 2008; Morris *et al.*, 2013).

Effect Size and Allele Frequency Estimation

Allele frequencies of the SNPs for both JL and GWAS were estimated using snpStats package in R (Clayton 2015). The proportion of phenotypic and genotypic variation explained by the JL and GWAS QTL were estimated using equations 6 and 7 (Utz *et al.*, 2000). The QTL additive effect sizes within and across families were both estimated as the difference between the mean of the two homozygous classes for each QTL divided by two. The additive effect size of each QTL identified in all models was estimated relative to RTx430 (Tian *et al.*, 2011). The proportion of phenotypic variation explained by each QTL was estimated by fitting a regression model with family and QTL as fixed terms;

$$y_{ijk} = \mu + \gamma_{\mathrm{i}} + \Phi_j + \varepsilon_{\mathrm{ijk}} \qquad [6]$$

where y_{ijk} is the phenotype, γ_{i} is the term for the NAM families, Φ_j is the term for QTL, and $\varepsilon_{\mathrm{ijk}}$ is the residual term. The sum of squares of QTLs divided by sum of squares total gave the proportion of variance explained by the detected QTL. In order to evaluate the within family variation explained by each QTL, a regression model was fit with terms for family and QTL nested within family as fixed effects (Würschum *et al.* 2011);

$$y_{ijkl} = \mu + \gamma_i + \omega_{jk} + \varepsilon_{ijkl} \qquad [7]$$

where y_{ijkl} is the phenotype, γ_i is the term for the NAM families, $\omega_{jk}$is the term for QTL nested within family, and ε_{ijkl} is the residual term. The sum of squares of QTLs (Family) divided by sum of squares total gave the within-family variance explained by the detected QTL (Würschum *et al.*, 2011).

Grass homologs search around identified loci

A set of *a priori* candidate genes (n = 39) associated with inflorescence morphology development were compiled from literature consisting of 24 maize genes, 10 rice genes, three sorghum genes, one foxtail millet gene, and one barley gene (supplementary files). Based on this candidate gene set, 297 sorghum homologs were found using Phytozome (Goodstein *et al.*, 2012). A custom R script was used to search for homologs within 150kb window both upstream and downstream of each association.

Cross-validation of NAM by Family

Cross-validation was performed using the ridge regression best linear unbiased prediction (rrBLUP) package in R (Endelman, 2011). First, the "leave-one-family-out" prediction approach was performed. This involves the removal of a family's genotypic and phenotypic data out of the whole NAM population and using remaining nine families to predict that particular family (NAM population minus family1 to predict family1). At each step of the analysis, a subagging approach that randomly samples data without replacement was used to sample 80% of each family. This step was repeated for all the 10 families in the NAM population for LBL and RL. The second type of analysis involved a five-fold cross-validation analysis for LBL, UBL, RL, and RD for 100 runs. The last approach was to perform family-by-family prediction similar to pairwise prediction between NAM families for LBL and RL. Prediction accuracy was estimated in each cycle as a correlation between predicted and observed phenotypic trait's value. Lastly, kinship relatedness between pairwise families and between the NAM and each family following the "leave-one-family-out" approach was estimated.

Results

Genome-wide polymorphism in the NAM population

A total of 116,405 SNPs were obtained after SNP calling, imputation, and filtering. SNP effect variant analysis identified a transition-transversion rate of 1. Missense, nonsense, and silent functional variants accounted for 71%, 6%, and 23% respectively, with a missense-silent variation ratio of 3.0. In addition, 2% exon variants, 15% intergenic variants, 65% intron variants, and 0.002% intragenic variants were identified. After filtering for 0.96 inbreeding coefficient, a total of 92,391 markers and 2220 RILs were identified. The number of RILs in each family range between 202 in Segaolane to 233 in SC265.

Variation in inflorescence morphology in the NAM population

Phenotypic variation distribution for the traits for each family showed that the mean trait value/performance of the RILs is greater than the performance of both parents in some families (Figures 2.1D). Significant genotypic differences were observed for traits (Table 2.3). The broad sense heritability estimates for the traits were high, ranging from 0.59 to 0.92. Based on trait-by-trait phenotypic correlations, RL and LBL had the highest correlation of 0.71 (*P*-value < 0.01). UBL and LBL both had a low positive correlation of 0.19 (*P*-value < 0.01) while RL had no correlation with UBL and RD (Figure 2.2).

QTL for inflorescence morphology

Significant QTL associations were observed for all traits (Figure 2.3 (A-D)). The within family and across NAM population effect of each QTL for both NJL and JL models were estimated relative to RTx430. Overall, JL identified more QTL than NJL. The within family additive effects of QTL in both models (NJL and JL) for all traits are listed in Table 2.4. The additive effects and proportion of phenotypic variation explained (PPVE) by JL QTL were greater than PPVE by NJL QTL (Table 2.5).

Comparison of NAM JL and diversity panel CMLM and GLM model results

The comparison of extent of LD decay in the SAP and NAM showed that LD decay rate was different between the SAP and NAM with faster decay in SAP compared to NAM (Figure 2.4A). Association mapping on the SAP for inflorescence lower branch length identified some of the genomic regions previously identified using the same data (Figures 2.3B). Most of the markers in the genome were associated with both LBL and RL in the naive GWAS model for SAP panel. However, in the CMLM model, only three associations were identified for LBL and no association for RL. The results of the SAP GLM and CMLM model were both plotted with the NAM LBL and RL JL results to see if there are co-localized associations between the mapping populations. Most of the significant associations (with *P*-values above permuted threshold) in the NAM were found at low significance levels below Bonferroni correction threshold in the SAP.

Genome-wide prediction of inflorescence morphology

For the leave-one-family-out approach, the mean prediction accuracy across all cycles for each trait was estimated and observed to range from 0.27 for SC1103 to 0.52 for SC971 for RL. For LBL, accuracy ranged from 0.21 in SC283 to 0.61 for SC1345. Relatively high prediction accuracies were observed for all the traits using the five-fold cross-validation approach. Prediction accuracies of 0.70, 0.65, 0.75, and 0.83 were observed for LBL, UBL, RL, and RD, respectively (Figures 2.5A, 2.5C, 2.10A, and 2.10B). Prediction accuracies were positively related to trait h^2 values ($r = 0.44$). For the family-by-family pairwise prediction accuracy, there was a positive relationship between pairwise family prediction and mean pairwise kinship relatedness for both LBL ($r = 0.19$ *P*-value < 0.05) and RL ($r = 0.1$ *P*-value < 0.05). In the “leave-one-family-out” approach, there was a non-significant positive relationship between prediction accuracy and kinship relatedness for RL ($r = 0.47$) and none for LBL ($r = 0.002$).

Discussion

Genetic basis of sorghum inflorescence morphology

This study using the NAM population provides insight about the genetic architecture of sorghum inflorescence morphology. Inflorescence morphology is controlled by numerous loci of minor effects (Table 2.5). Few major and many minor effect loci were also found to underlie rice and maize inflorescence traits (Crowell *et al.*, 2016; Wu *et al.*, 2016). Additionally, the variable effect sizes of the underlying loci suggest multiple gene changes may be required to produce adaptive phenotypic change in sorghum inflorescence (Lauter & Doebley, 2001). QTL detected in the whole NAM population using JL showed contrasting allele substitution effects within families. Some families had positive effects while others had negative effects for the same QTL (Figures 2.6–9). This differing effects direction can be attributed to possible epistatic interactions of the QTL with other loci within families.

Based on comparison with earlier mapping studies in sorghum, some QTL identified in this study were previously known while others are novel. Among the previously known QTL is a pleiotropic QTL, qSbRL7.59, associated with both LBL and RL, centered on the intragenic region of a *YUCCA5* homolog (flavin monooxygenase gene) and 69.9 kb away from the sorghum height gene (*Dw3*, a phosphoglycoprotein gene, Sobic.007G163800) (Figure 2.3A and 2.3C). Previous linkage mapping studies identified association around this same *Dw3* region for a QTL that increased rachis length and primary branch length (Brown *et al.*, 2006; Shehzad & Okuno, 2015) and *YUCCA5* was proposed as a candidate for the gene underlying branch length variation (Brown *et al.*, 2008). Some of the QTL in this study were found to be novel associations when compared to associations from reanalyzed data from a previously published GWAS (Morris *et al.*, 2013) in sorghum using GLM (naive model) and CMLM (full model) (Figure 2.4B). This could be due to the power of the NAM in reducing the effects of ancestral population structure on mapping.

About 46% (58 of 127) of the inflorescence QTL identified in this study co-localized with sorghum homologs of maize and rice inflorescence genes within a 150kb window (Table 2.6). A QTL (SbInf_03.4750) for UBL was found about 38 kb from the sorghum ortholog (Sobic.003G052900) of maize *ramosa2* (*ra2*) (Figure 2.3B). For LBL, a QTL (qSbLBL9.49) was found in the intragenic region of Sobic.009G142200 (*No Apical Meristem* gene) involved in floral organ identity and development. For RL, a QTL (qSbRL3.57) was found outside the 150kb window, about 160 kb from the sorghum ortholog (Sobic.003G236900) of the maize *YUCCA2* gene (*spi1*) (Figure 2.3C). Mutations in *YUCCA* gene(s) led to drastic reduction in inflorescence rachis length and branch length in maize and rice and general deformity in Arabidopsis inflorescence (Cheng *et al.*, 2006; Yamamoto *et al.*, 2007; Kim *et al.*, 2007; Gallavotti *et al.*, 2008). In general, there was no colocalization between QTL for LBL and UBL except qSbUBL3.44 and qSbUBL3.47 which were ~300 kb from each other falling outside the LD range (150 kb) in this population. This suggests that LBL and UBL are under different genetic control. The fact that some QTL are far from *a priori* candidate genes may in part reflect lower mapping resolution in the NAM population compared to the SAP due to slower LD decay in the NAM population (Figure 2.4A).

Genomic enabled breeding of inflorescence morphology

The inflorescence traits evaluated in this study had high heritability, which signify the presence high genetic variation for increased selection gain. Given the high prediction accuracies observed for the traits (from $r = 0.65$ to $r = 0.83$; Figures 2.5A and 2.5C); thus, genomic prediction is possible for inflorescence morphology in sorghum. As this study was based on RTx430-derived NAM families, prediction in related or breeding populations of similar pedigree (e.g., from Texas A&M breeding lines) will also be beneficial for high accuracy genome-wide predictions in sorghum. The variation in prediction accuracies obtained in the "leave-one-family-out" approach can be a reflection of the differences in the genetic diversity captured by the NAM for each of the five sorghum botanical races. Between family predictions in the maize NAM also showed varied prediction accuracies (Peiffer *et al.*, 2014). Only four of the five racial groups were represented in the NAM population with disproportionate number of families representing each race. Increasing the number of NAM founders with sufficient racial representation will be beneficial in capturing more genetic diversity underlying sorghum inflorescence for higher predictive power. However, prediction accuracies using family-to-family pairwise prediction approaches did not show a strong positive relationship ($r < 0.2$) with kinship relatedness for RL and LBL.

Sorghum breeding programs often must cross parental lines with contrasting panicle morphologies, especially to transfer traits from donor to recipient genetic backgrounds. Unfortunately, recovering recipient panicle morphology by backcross and/or intercross is slow, because phenotypic selection for inflorescence morphology can only be done close to harvest time, well after the window for pollination. Therefore, QTL markers from this study can be employed to facilitate recovery of desired inflorescence morphology via marker-assisted recurrent selection. The pleiotropy of qSbRL7.59 with both LBL and RL and their high phenotypic correlation ($r = 0.71$, P-value < 0.01) can be beneficial for simultaneous selection of both traits, for example, in the recovery of large panicle from a cross between small and large panicle sorghum lines.

Conclusions

Sorghum NAM demonstrated its power in dissecting population structured adaptive traits in this study by validation of previously reported QTL and identification of novel ones. Sorghum inflorescence is controlled by loci of minor effects. Some, but not all, of the *a priori* candidate genes were associated with variation in inflorescence morphology. Most of the inflorescence *a priori* genes were identified in mutation studies; thus variants identified by such approach may not be reflective of natural populations, where deleterious large-effect variants will be purged by natural selection. There is also likely more genetic diversity in inflorescence morphology to be discovered, since the 11 NAM founder parents only captured about 70% of the U.S. sorghum association panel (Bouchet *et al.*, 2017). Therefore, increasing the number of the founders in the NAM population (i.e. adding new families) will be beneficial for both increased recombination and diversity. Although this may increase phenotyping burden, the use of high-throughput phenotyping platforms could overcome this challenge (Crowell *et al.*, 2016). Furthermore, since the NAM founders originated from diverse agroclimatic zones, QTL identified in this study should be transferable for MAS across breeding programs in various climatic zones. In breeding programs targeting smallholder-farming systems, my QTL offers an opportunity to recover farmer-preferred inflorescence traits. This will be possible by using them as markers to recover locally preferred inflorescence morphology when introgressing traits into local genetic backgrounds.

Tables and Figures

Table 2.1: The sorghum NAM founders, their origin and number of RILs used for analysis from each diverse founder derived family.

Founder	Origin	Founder Type	RILs
RTX430	Texas A & M University	Common Parent	-
P898012	Purdue University	Diverse Founder	213
Ajabsido	Sudan	Diverse Founder	214
Macia	ICRISAT	Diverse Founder	231
SC1103	Nigeria	Diverse Founder	231
SC1345	Mali	Diverse Founder	231
SC265	Burkina Faso	Diverse Founder	232
SC283	Tanzania	Diverse Founder	223
SC35	Ethiopia	Diverse Founder	208
SC971	Puerto Rico, United States	Diverse Founder	233
Segaolane	Botswana	Diverse Founder	204

Table 2.2: Location, climate, year, precipitation (rainfall October of previous year to October of current year) and environmental code information of field experiments for the nested association mapping population.

Location	Climate	Year	Precipitation (mm) Oct – Oct*	Environment Code
Manhattan, KS	Humid Continental	2014	698	MN14
Hays, KS	Semi-Arid	2014 (Upland)	639	HA14
Manhattan, KS	Humid Continental	2015	998	MN15
Hays, KS	Semi-Arid	2015 (Bottomland)	513	HI15
Hays, KS	Semi-Arid	2015 (Upland)	513	HD15

* National Oceanic and Atmospheric Administration, U.S. Department of Commerce.

Table 2.3: Mean, range, and broad sense heritability (H^2) for lower branch length (LBL), upper branch length (UBL), rachis length (RL), and rachis diameter (RD).

Trait	Mean (mm)	Range (mm)	H^2
LBL***	82	267 – 176	0.86
UBL*	48	7 – 170	0.85
RL***	274	111 – 465	0.92
RD***	8.3	3.8 - 13.5	0.59

*, **, *** Significant genotypic differences at 0.05, 0.01 and 0.001 respectively.

Table 2.4: Within family additive effect size (AES) for QTL identified using joint linkage mapping with marker nested within family (NJL) and joint linkage with no nesting (JL).

Trait	Model	Nos. of QTL	Range of AES (mm)
LBL	NJL	14	-30 to16
LBL	JL	21	-26 to 19
UBL	NJL	1	- 28.0 to 0
UBL	JL	17	-33 to 5
RL	NJL	16	-44 to 51.9
RL	JL	22	-49.5 to 51.9
RD	NJL	9	-2.0 to 0.9
RD	JL	21	-2.4 to 0.6

Table 2.5: Across population (whole NAM) additive effect size (AES) and proportion of phenotypic variation explained for QTL identified using joint linkage mapping with marker nested within family (NJL) and joint linkage with no nesting (JL).

Trait	Model	Range of AES (mm)	Range of PPVE (%)
LBL	NJL	-4 to 2	0.1 to 2.0
LBL	JL	-12 to 8	0.6 to 5.0
UBL	NJL	-4	3.0
UBL	JL	-11 to 2	0.6 to 4.0
RL	NJL	-10 to 12	0.1 to 3.0
RL	JL	-20 to 20	0.6 to 3.0
RD	NJL	-0.4 to 0.2	0.1 to 1.0
RD	JL	-0.9 to 1.5	0.2 to 1.0

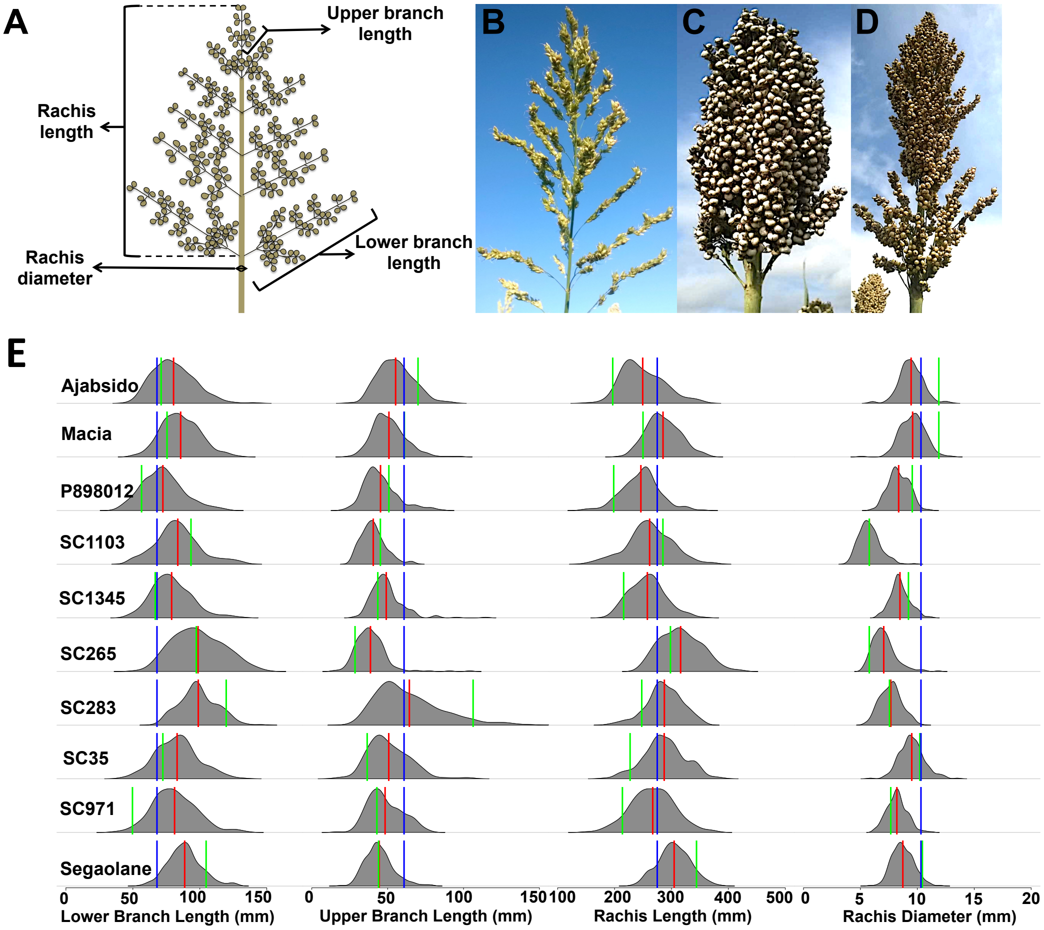

Figure 2.1: Phenotypic description and distribution of sorghum inflorescence morphology.

(A) Diagram of sorghum inflorescence traits evaluated. (B) Open inflorescence morphology represented by SC1103 parent. (C) Compact inflorescence morphology represented by Ajabsido parent. (D) Semi-compact inflorescence morphology as represented by RTx430 the common parent. (E) Phenotypic distribution of line means for inflorescence traits. The blue lines are the trait value for the common parent (RTx430), the green lines are the mean trait values for each of the other parent, and the red line is the trait mean within each family.

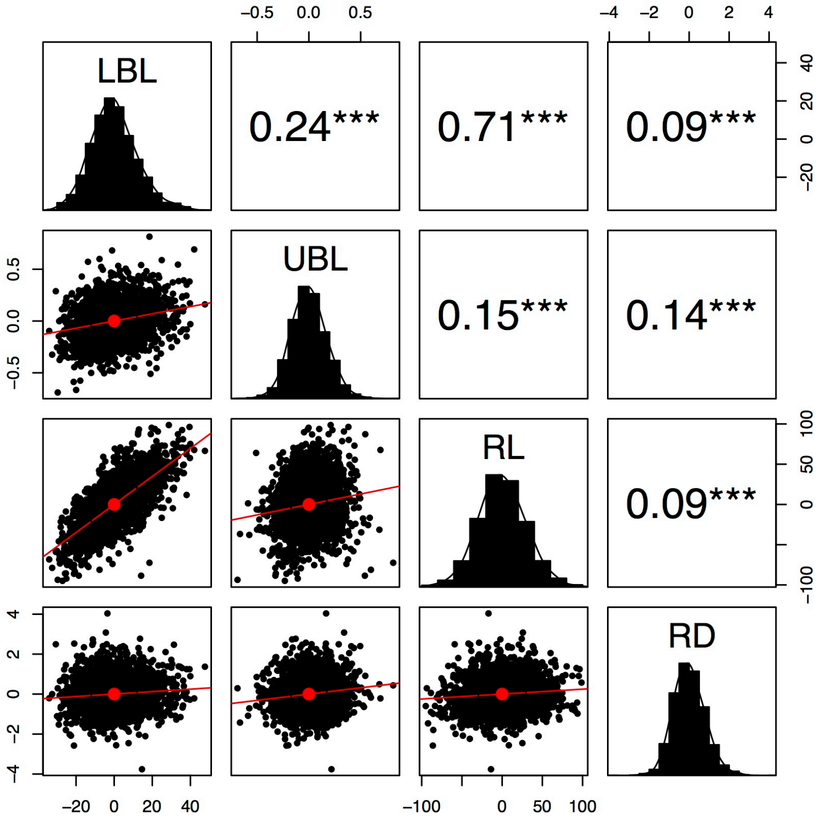

Figure 2.2: Pairwise correlation between traits.

Pearson correlation between lower branch length (LBL), upper branch length (UBL), rachis length (RL), and rachis diameter (RD) significant at 0.05, 0.01 and 0.001 (*, **, and ***).

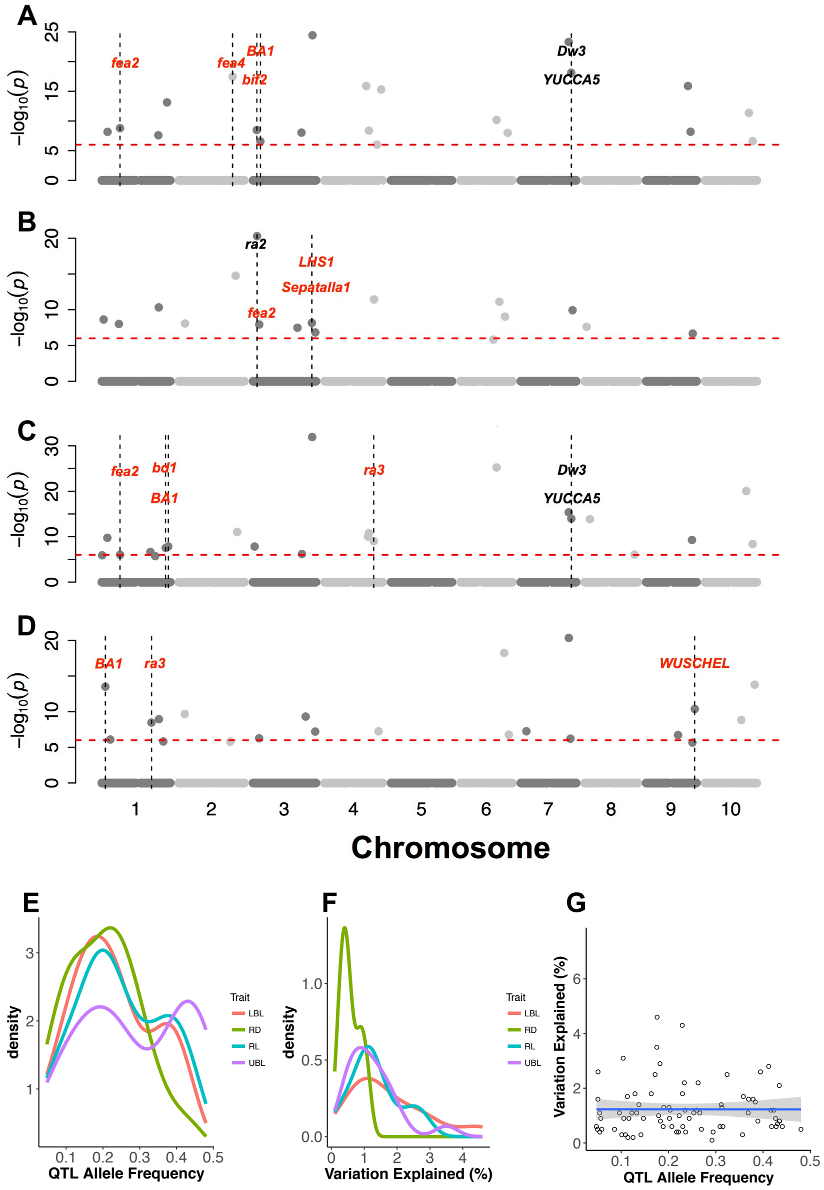

Figure 2.3: QTL mapping for inflorescence morphology using joint linkage model.

Genomic location of associations with (A) lower branch length, (B) upper branch length, (C) rachis length, and (D) rachis diameter. The dashed red lines are the Bonferroni significance threshold (*P*-value < 0.05) estimated from 100 permutations. *A priori* candidate genes that colocalize with QTL within 150 kb are noted as follows. Black text indicates putative sorghum orthologs of *a priori* candidate genes while red text indicates paralogs. (E-F) Density plots showing the distribution of QTL allele frequency and phenotypic variation explained for each trait. (G) Plot showing the relationship between QTL allele frequency and phenotypic variation explained.

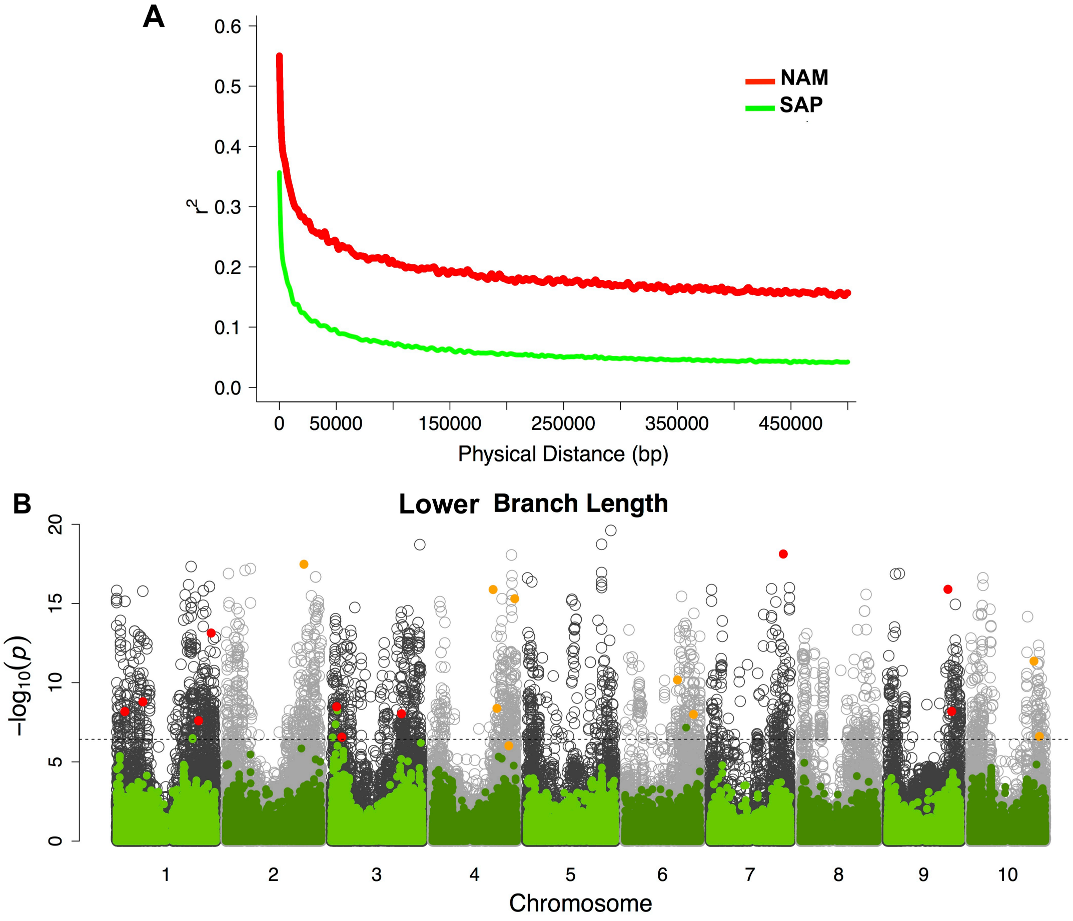

Figure 2.4: Comparison of sorghum diversity/association panel (SAP) and nested association mapping (NAM) populations based on linkage disequilibrium decay and genome-wide association mapping.

(A) Linkage disequilibrium decay curve color coded for sorghum association panel (SAP) and nested association mapping (NAM) populations and (B) Manhattan plot for the comparison of genome-wide association approaches for lower branch length in using generalized liner model (GLM) in shades of gray, compressed mixed linear model (CMLM) in shades of green, and NAM joint linkage (JL) model in red and orange.

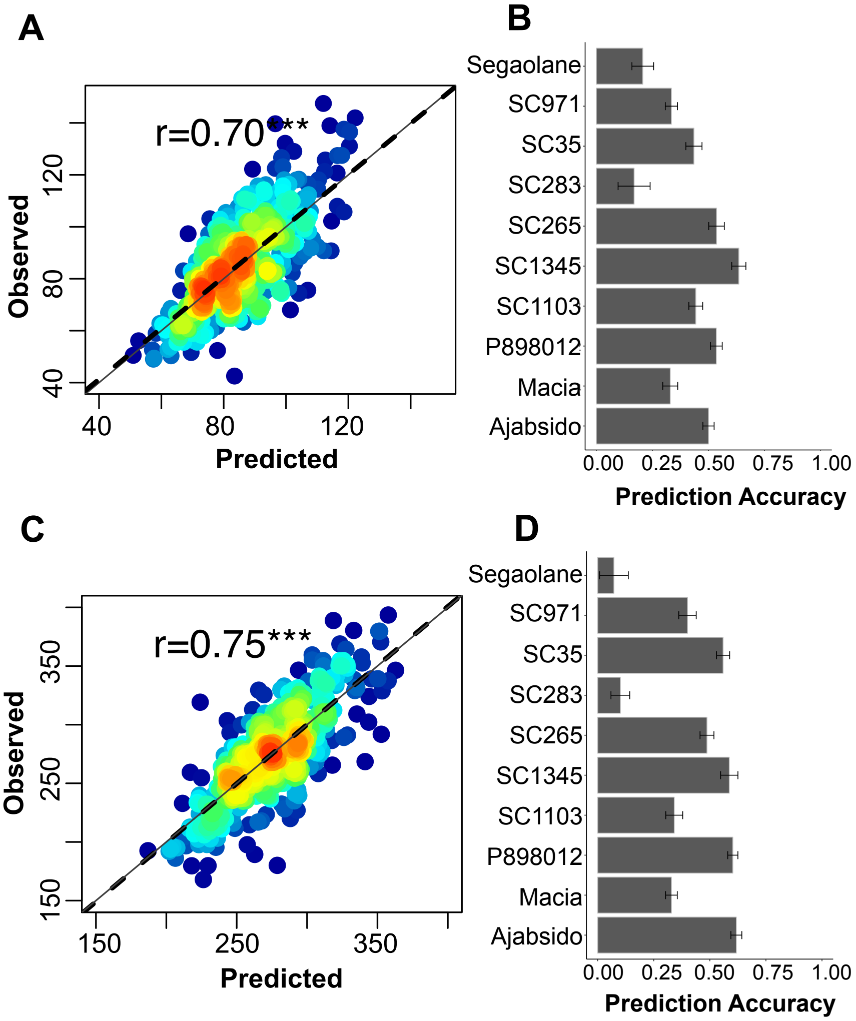

Figure 2.5: Prediction accuracies for five-fold cross-validation and leave-one-family-out subagging approaches for lower branch length and rachis length.

(A) The prediction accuracy as correlation between observed phenotypic values and predicted phenotypic values for lower branch length cross-validation. The color intensity shows the density of data points in each region (from blue to red means few data points to highly dense data points). (B) Prediction accuracies using leave-one-out approach for each family for lower branch length. (C) The prediction accuracy as correlation between observed phenotypic values and predicted phenotypic values for rachis length cross-validation. The color intensity shows the density of data points in each region (from blue to red means few data points to highly dense data points). (D) Prediction accuracies using leave-one-out approach for each family for rachis length.

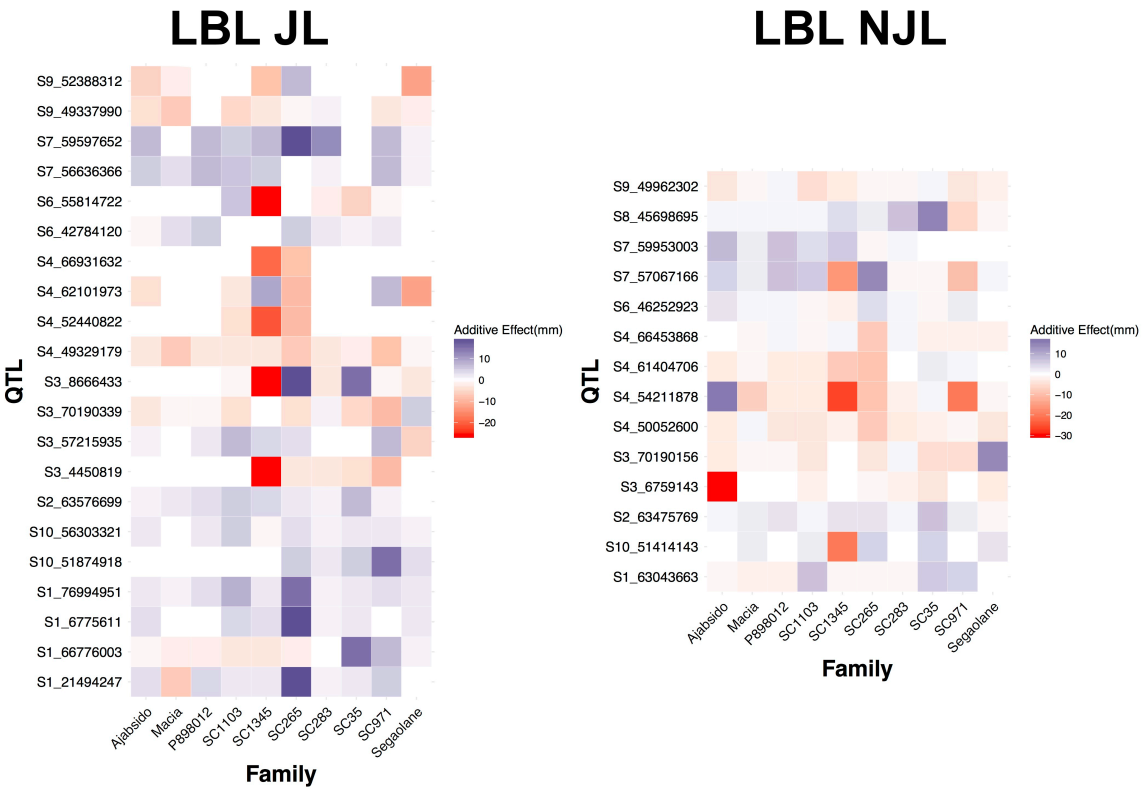

Figure 2.6: Joint linkage and nested joint linkage (NJL) additive effects heatmap for lower branch length (LBL).

The red boxes indicate families where the additive effects of the QTL are negative. While the blue boxes indicate families where the additive effects of the QTL are positive.

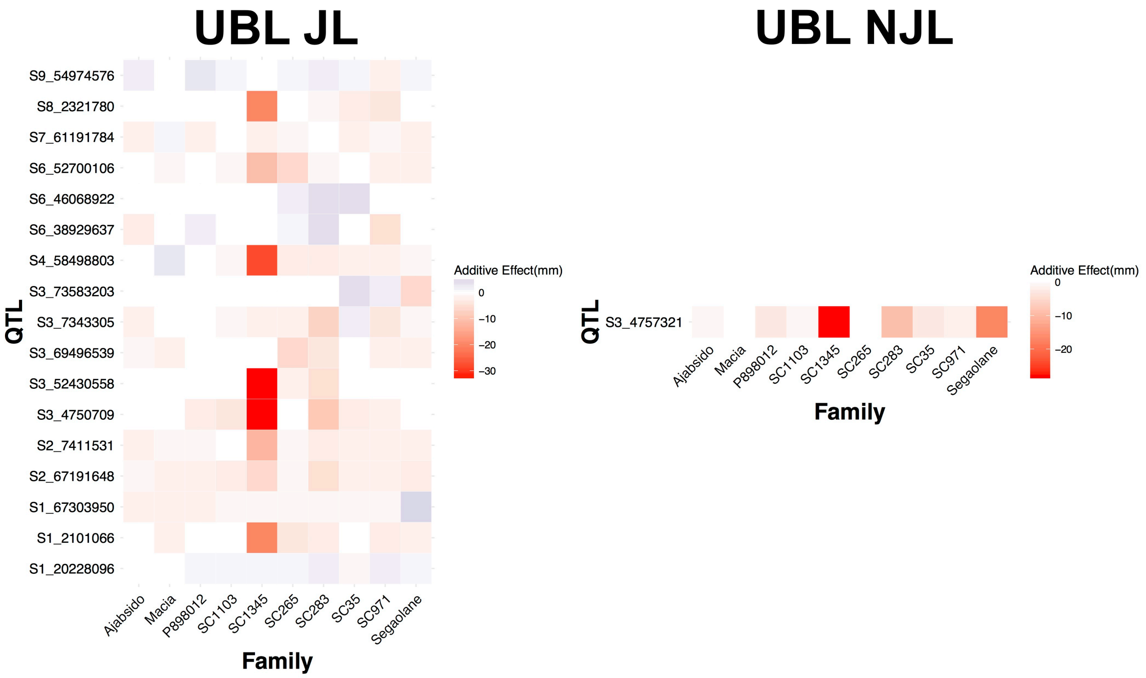

Figure 2.7: Joint linkage and nested joint linkage (NJL) additive effects heatmap for upper branch length (UBL).

The red boxes indicate families where the additive effects of the QTL are negative. While the blue boxes indicate families where the additive effects of the QTL are positive.

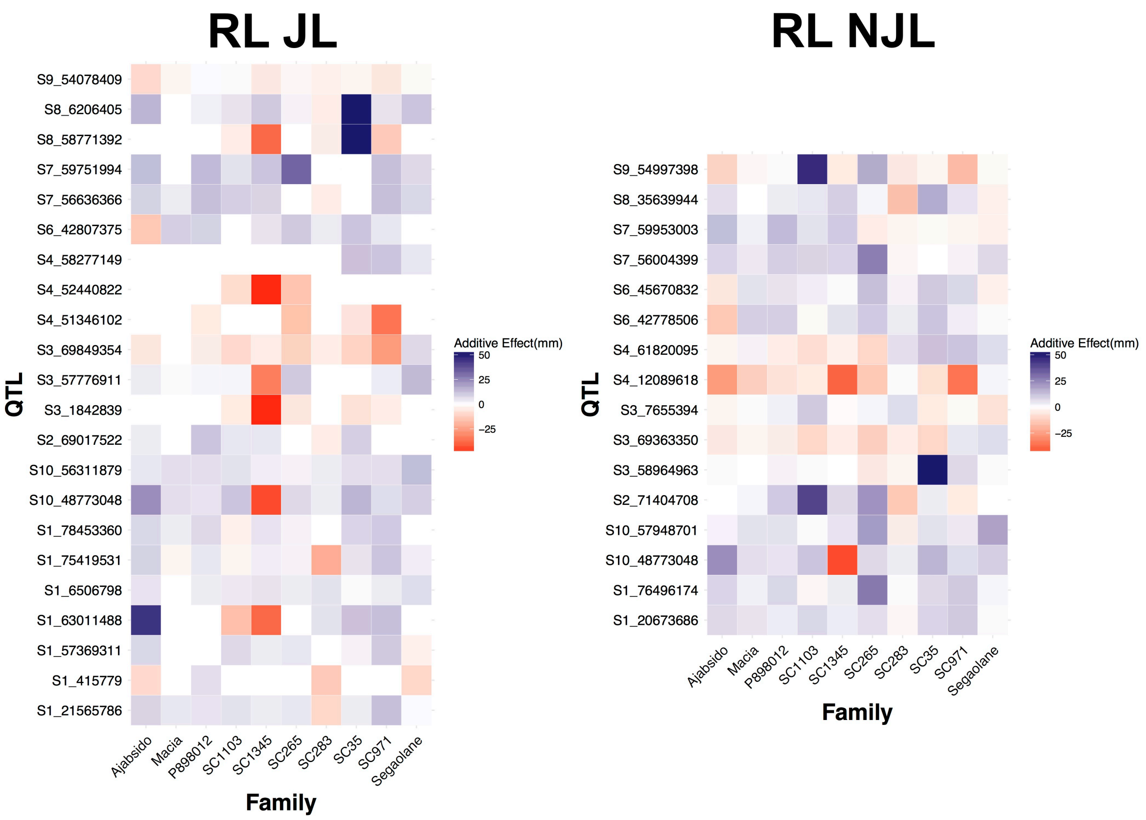

Figure 2.8: Joint linkage and nested joint linkage (NJL) additive effects heatmap for rachis length (RL).

The red boxes indicate families where the additive effects of the QTL are negative. While the blue boxes indicate families where the additive effects of the QTL are positive.

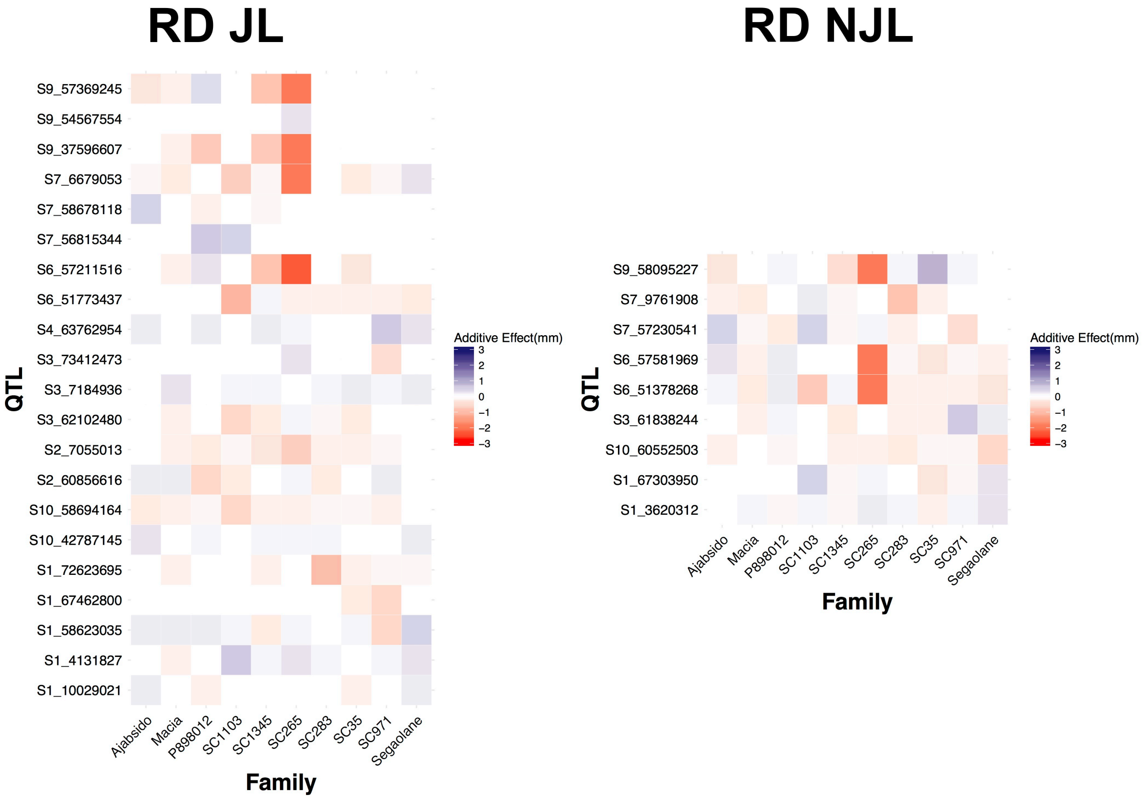

Figure 2.9: Joint linkage and nested joint linkage (NJL) additive effects heatmap for rachis diameter (RD).

The red boxes indicate families where the additive effects of the QTL are negative. While the blue boxes indicate families where the additive effects of the QTL are positive.

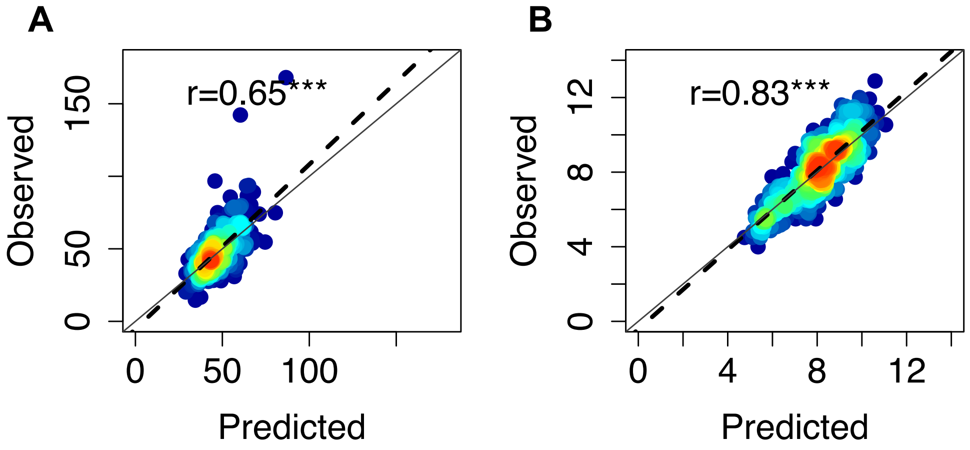

Figure 2.10: Prediction accuracies for five-fold cross-validation.

Prediction accuracy plots showing correlation between observed phenotypic values and predicted phenotypic values for (A) upper branch length and (B) rachis diameter. The color intensity shows the density of data points in each region (from blue to red means few data points to highly dense data points).

References

Bai F, Reinheimer R, Durantini D, Kellogg EA, Schmidt RJ. **2012**. TCP transcription factor, BRANCH ANGLE DEFECTIVE 1 (BAD1), is required for normal tassel branch angle formation in maize. *Proceedings of the National Academy of Sciences of the United States of America* **109**: 12225–30.

Bastide H, Lange JD, Lack JB, Yassin A, Pool JE. **2016**. A Variable Genetic Architecture of Melanic Evolution in Drosophila melanogaster. *Genetics* **204**: 1307–1319.

Bates D, Maechler M, Bolker B, Walker S, Christensen RHB, Singmann H, Dai B, Eigen C. **2017**. Fitting linear mixed-effects models using lme4. *Journal of statistical software* **67**: 1–113.

BGI-shenzhen. **2017**. BGI-shenzhen/PopLDdecay - Libraries.io. https://github.com/BGI-shenzhen/PopLDdecay

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.

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.

Brown PJ, Rooney WL, Franks C, Kresovich S. **2008**. Efficient Mapping of Plant Height Quantitative Trait Loci in a Sorghum Association Population With Introgressed Dwarfing Genes. *Genetics* **180**: 629–637.

Brown PJ, Upadyayula N, Mahone GS, Tian F, Bradbury PJ, Myles S, Holland JB, Flint-Garcia S, McMullen MD, Buckler ES, *et al.* 2011. Distinct genetic architectures for male and female inflorescence traits of maize. *PLoS Genetics* **7**: e1002383.

Browning BL, Browning SR. **2013**. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**: 459–71.

Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, Ersoz E, Flint-Garcia S, Garcia A, Glaubitz JC, *et al.* 2009. The Genetic Architecture of Maize. *Science* **325**: 714–718.

Casa AM, Pressoir G, Brown PJ, Mitchell SE, Rooney WL, Tuinstra MR, Franks CD, Kresovich S. **2008**. Community resources and strategies for association mapping in Sorghum. *Crop Science* **48**: 30–40.

Cheng Y, Dai X, Zhao Y. **2006**. Auxin biosynthesis by the YUCCA flavin monooxygenases controls the formation of floral organs and vascular tissues in Arabidopsis. *Genes Dev* **20**: 1790–1799.

Chuck G, Muszynski M, Kellogg E, Hake S, Schmidt RJ, McSteen P, Laudencia-Chingcuanco D, Colasanti J, Cheng PC, Greyson RI, *et al.* 2002. The control of spikelet meristem identity by the branched silkless1 gene in maize. *Science (New York, N.Y.)* **298**: 1238–41.

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: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**: 80–92.

Cresko WA, Amores A, Wilson C, Murphy J, Currey M, Phillips P, Bell MA, Kimmel CB, Postlethwait JH. **2004**. Parallel genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback populations. *Proceedings of the National Academy of Sciences of the United States of America* **101**: 6050–5.

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**: 10527.

Endelman JB. **2011**. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome Journal* **4**: 250.

Gallavotti A, Barazesh S, Malcomber S, Hall D, Jackson D, Schmidt RJ, McSteen P. **2008**. sparse inflorescence1 encodes a monocot-specific YUCCA-like gene required for vegetative and reproductive development in maize. *Proceedings of the National Academy of Sciences of the United States of America* **105**: 15196–15201.

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 (NA Tinker, Ed.). *PLoS ONE* **9**: e90346.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, *et al.* 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic acids research* **40**: D1178-86.

Huang P, Jiang H, Zhu C, Barry K, Jenkins J, Sandor L, Schmutz J, Box MS, Kellogg EA, Brutnell TP. **2017**. Sparse panicle1 is required for inflorescence development in Setaria viridis and maize. *Nature Plants* **3**: 17054.

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.

Hung H-Y, Browne C, Guill K, Coles N, Eller M, Garcia A, Lepak N, Melia-Hancock S, Oropeza-Rosas M, Salvo S, *et al.* 2012. The relationship between parental genetic or phenotypic divergence and progeny variation in the maize nested association mapping population. *Heredity* **108**: 490–9.

Ishii T, Numaguchi K, Miura K, Yoshida K, Thanh PT, Htun TM, Yamasaki M, Komeda N, Matsumoto T, Terauchi R, *et al.* 2013. OsLG1 regulates a closed panicle trait in domesticated rice. *Nature genetics* **45**: 462–5, 465–2.

Jiang G, Xiang Y, Zhao J, Yin D, Zhao X, Zhu L, Zhai W. **2014**. Regulation of inflorescence branch development in rice through a novel pathway involving the pentatricopeptide repeat protein sped1-D. *Genetics* **197**: 1395–407.

Kellogg E. **2007**. Floral displays: genetic control of grass inflorescences. *Current Opinion in Plant Biology* **10**: 26–31.

Kim JI, Sharkhuu A, Jin JB, Li P, Jeong JC, Baek D, Lee SY, Blakeslee JJ, Murphy AS, Bohnert HJ, *et al.* 2007. yucca6, a dominant mutation in Arabidopsis, affects auxin accumulation and auxin-related phenotypes. *Plant physiology* **145**: 722–35.

Lauter N, Doebley J. **2001**. Genetic Variation for Phenotypically Invariant Traits Detected in Teosinte: Implications for the Evolution of Novel Forms. *Genetics* **160**: 333–342.

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.

Maurer A, Draba V, Jiang Y, Schnaithmann F, Sharma R, Schumann E, Kilian B, Reif JC, Pillen K. **2015**. Modelling the genetic architecture of flowering time control in barley through nested association mapping. *BMC genomics* **16**: 290.

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**: 453–8.

Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, Costich DE, Buckler ES. **2009**. Association Mapping: Critical Considerations Shift from Genotyping to Experimental Design. *The Plant Cell* **21**: 2194–2202.

National Research Council. **1996**. Lost Crops of Africa. In: BOSTID (Board on Science and Technology for International Development). National Academy Press, 383.

Pann K, Hmon W, Tariq S, Okuno K, Witt Hmon KP, Shehzad ÁT, Okuno ÁK, Shehzad T. **2014**. QTLs underlying inflorescence architecture in sorghum. *Genet Resour Crop Evol* **61**: 1545–1564.

Paterson AH. **2013**. Genomics of the Saccharinae. *Genomics of the Saccharinae*: 1–567.

Peiffer JA, Romay MC, Gore MA, Flint-Garcia SA, Zhang Z, Millard MJ, Gardner CAC, McMullen MD, Holland JB, Bradbury PJ, *et al.* 2014. The genetic architecture of maize height. *Genetics* **196**: 1337–1356.

Poland JA, Bradbury PJ, Buckler ES, Nelson RJ, Major Goodman by M. **2011**. Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. *Proceedings of the National Academy of Sciences* **108**: 6893–6898.

RAKSHIT S, RAKSHIT A, PATIL J V. 2012. Multiparent intercross populations in analysis of quantitative traits. *Journal of Genetics* **91**: 111–117.

Satoh-Nagasawa N, Nagasawa N, Malcomber S, Sakai H, Jackson D. **2006**. A trehalose metabolic enzyme controls inflorescence architecture in maize. *Nature* **441**: 227–230.

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.

Shehzad T, Okuno K. **2015**. QTL mapping for yield and yield-contributing traits in sorghum (Sorghum bicolor (L.) Moench) with genome-based SSR markers. *Euphytica* **203**: 17–31.

Tanaka W, Pautler M, Jackson D, Hirano HY. **2013**. Grass meristems II: Inflorescence architecture, flower development and meristem fate. *Plant and Cell Physiology* **54**: 313–324.

Tian F, Bradbury PJ, Brown PJ, Hung H, Sun Q, Flint-Garcia S, Rocheford TR, McMullen MD, Holland JB, Buckler ES. **2011**. Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nature genetics* **43**: 159–62.

Utz HF, Melchinger AE, Schön CC. **2000**. Bias and sampling error of the estimated proportion of genotypic variance explained by quantitative trait loci determined from experimental data in maize using cross validation and validation with independent samples. *Genetics* **154**: 1839–1849.

Wallace JG, Larsson SJ, Buckler ES. **2014**. Entering the second century of maize quantitative genetics. *Heredity* **112**: 30–8.

Witt Hmon KP, Shehzad T, Okuno K. **2013**. Variation in inflorescence architecture associated with yield components in a sorghum germplasm. *Plant Genetic Resources* **11**: 258–265.

Wu X, Li Y, Shi Y, Song Y, Zhang D, Li C, Buckler ES, Li Y, Zhang Z, Wang T. **2016**. Joint-linkage mapping and GWAS reveal extensive genetic loci that regulate male inflorescence size in maize. *Plant biotechnology journal* **14**: 1551–1562.

Würschum T, Liu W, Gowda M, Maurer H, Fischer S, Schechert A, Reif J. **2011**. Comparison of biometrical models for joint linkage association mapping. *Heredity* **108**: 332–340.

Yamamoto Y, Kamiya N, Morinaka Y, Matsuoka M, Sazuka T. **2007**. Auxin biosynthesis by the YUCCA genes in rice. *Plant physiology* **143**: 1362–1371.

Yu J, Holland JB, McMullen MD, Buckler ES. **2008**. Genetic design and statistical power of nested association mapping in maize. *Genetics* **178**: 539–51.

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, *et al.* **2006**. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* **38**: 203–208.

Chapter 3 - Joint Linkage Mapping of Vegetative Traits in Sorghum

Abstract

Crop improvement strategies for increased yield often involve modification of plant morphology. In sorghum, the genetic basis of plant height has been well described, but traits such as stem diameter, leaf erectness, and leaf width are not well understood. In this study, the sorghum-nested association mapping (NAM) population was used to characterize the genetic architecture of leaf erectness, leaf width, and stem diameter. About 2200 recombinant inbred lines were phenotyped in multiple environments. The obtained phenotypic data were used to perform joint linkage mapping using ~93,000 markers. Minor effects quantitative traits loci (QTL) were found to underlie marker-trait associations as the explained very small proportion of the phenotypic variation (< 10 %). A pleiotropic QTL was found to be associated with plant height, stem diameter, and leaf erectness. Furthermore, identified QTL co-localized with sorghum homologs of developmental genes in other grasses. Our results provide insights into the genetic basis of leaf erectness and stem diameter in sorghum. The QTL identified could facilitate molecular characterization of the genes underlying vegetative growth and development, and genomic-enabled breeding for improved plant architecture.

Introduction

Modification of plant morphology for crop improvement has contributed to global agricultural productivity during the last century (Khush, 2001; Duvick, 2005). This approach known as ideotype breeding involves the creation of a model plant with characteristics that facilitate efficient photosynthesis, growth, and yield (grain and biomass) (Donald, 1968; Hammer *et al.*, 2009). Ideotype breeding also contributed to increased yield potential in maize, rice, and wheat during the "Green Revolution" (Khush, 2001). The Green Revolution ideotype in cereals includes reduced height, erect leaves, thick stalks, large and semi-compact inflorescence (ear, panicle, or head). Leaf erectness is defined the angle between the soil level and the leaf midrib, where more erect leaves have greater angle (Tian *et al.*, 2011). Erect leaves are thought to have contributed directly or indirectly to increased grain yield in U.S maize through adaptation of hybrids to high planting densities (Duvick, 2005; Hammer *et al.*, 2009). In sorghum, increase leaf erectness was shown to improve photosynthetic efficiency and thermal stress reduction by dispersing solar radiation from upper to lower parts of the canopy (Truong *et al.*, 2015). Wide leaves may also facilitate efficient solar radiation capture for plant productivity. Thick stems may increase stalk strength and improve standability (lodging resistance) for combine harvesting (Kashiwagi *et al.*, 2008). A wide genetic variation exists for these ideotype traits in cereals that can be further utilized in crop improvement (Khush, 2001). However, the genetic basis of these agronomically important traits is not well characterized in sorghum.

Characterization of trait genetic architecture is made possible by quantitative trait loci (QTL) mapping (Morrell *et al.*, 2011). Previous and current crop trait dissection involved the use of bi-parental populations for mapping. But the power and reliability of dissection using biparental mapping are often undermined by small population size, reduced genetic diversity, and non-transferability of QTL to other backgrounds. Association mapping panels exploit wide genetic diversity and provide high mapping resolution for genetic dissection of traits. However, the effectiveness of association mapping is limited by population structure that causes spurious and synthetic associations (Myles *et al.*, 2009; Korte & Farlow, 2013). Multi-parental mapping like the nested association mapping (NAM) and multiple advanced generation intercrosses (MAGIC) reduce the confounding effect of historical population structure and phenotypic variation through crosses between founder lines and common founder in NAM and among founder lines in MAGIC (Myles *et al.*, 2009; Korte & Farlow, 2013). The balanced allele frequencies in the NAM population also contribute to its high mapping power. Thus, it is an efficient approach for characterizing the genetic basis of complex agronomic traits (Bouchet *et al.*, 2017).

Given that agronomic traits have shaped by selection, considering the evolutionary history of selection provides insights on genetic architecture of agronomic traits. Evidences of the impact of selection on the evolution history of genetic architecture have been shown in fish (Cresko *et al.*, 2004), dogs (Boyko *et al.*, 2010), chicken (Carlborg *et al.*, 2006), maize (Doebley, 2004; Brown *et al.*, 2011), and rice (McCouch *et al.*, 2004; Ishii *et al.*, 2013). Traits that have undergone recent selection have been found to be under the control of large effect loci in maize (Brown *et al.*, 2011), rice (Ishii *et al.*, 2013), and sorghum (Bouchet *et al.*, 2017).

Sorghum is an important staple food crop in semi-arid and arid regions due to its resilience to harsh environmental conditions. Its cultivation in temperate climate during the last century was made possible by the conversion of tall photoperiod-sensitive tropical lines to dwarf photoperiod-insensitive high yielding lines with reduced height. This improved sorghum ideotype for temperate climates made sorghum commercial cultivation successful in the United States (Klein *et al.*, 2008). Selection for reduced height led to indirect selection for increased erectness of sorghum leaves and stem diameter that made high-density cultivation more feasible. Genetic architecture of flowering time and height has been characterized in sorghum to a great extent unlike leaf morphology and stem diameter. The major genes underlying height in sorghum are *Dw1*, *Dw2,* and *Dw3*. *Dw3* has been shown to have a pleiotropic effect on leaf erectness and inflorescence in the lower primary branch in sorghum (Brown *et al.*, 2008). Flowering time in sorghum is under the control of maturity gene loci (*ma1-ma6*). In maize, the genetic architecture of leaf erectness and leaf width has been well characterized and found to be controlled by small effect loci (Tian *et al.*, 2011; Strable *et al.*, 2017).

In cereals, the shoot apical meristem (SAM) determines the above ground architecture of the plant by maintaining pluripotent cells and forming lateral organs and stems which both determine the various plant architectures observed in nature (Wang & Li, 2008). Genes such as *LA1* (*LAZY*), *TAC1* (*Tillering Angle Control 1*) (Wang and Li 2008), *Narrow leaf1* (Qi *et al.*, 2008), *LC1 (LEAF INCLINATION1)* (Zhao *et al.*, 2013), *lg1* (*liguless1*), *lg2* (*liguless2*), *lg3* (*liguless3*), and *lg4* (*liguless4*), and *YABBY* (Tian *et al.*, 2011; Li *et al.*, 2015; Strable *et al.*, 2017) play essential roles in lateral organ branching, leaf formation, tiller angle, and leaf angle regulation. In addition, growth-promoting compounds as gibberellins and brassinosteroids also regulate plant architecture (Wang & Li, 2008). Currently, the genetic basis of leaf morphology and stem diameter in sorghum is poorly understood. In this study, the genetic basis of sorghum leaf erectness, leaf width, and stem diameter was characterized using the sorghum NAM population. This population comprised of approximately 2200 RILs, generated from a cross between 10 diverse founder lines with the common parent. The study objectives were (i) to characterize the genetic architecture (number, effect size, and allele frequencies of the underlying QTL) and (ii) to identify genes underlying QTL associated with the above-mentioned traits.

Materials and Methods

Plant materials and phenotypic evaluations

The sorghum NAM population was previously described (Paterson, 2013; Bouchet *et al.*, 2017). It consists of 2500 recombinant inbred lines (RILs) derived from a cross between a common founder RTx430 and 10 other diverse founders (Table 2.1). This population was phenotyped in multiple environments (Table 2.2) for leaf erectness, leaf width, stem diameter, flowering time and height. The leaf erectness was measured as the angle between the soil surface (0°) and the pre-flag leaf midrib using a barcode reader and barcode protractor (Figure 2.1). Leaf erectness was measured from 3 plants per plot. The leaf width was measured as the width of the leaf at the widest point on both pre-flag leaf and the fourth leaf from the flag leaf on three plants per plot (Figure 2.1). Measurements were taken using a barcode ruler. Stem diameter was measured as the diameter of the stem at the second Internode above the ground surface on three plants per plot in millimeters using a digital caliper and a barcode reader. Flowering time was scored as the number of days from planting to the day in which 50% of the individuals in a plot are in anthesis.

Phenotypic data analysis

Analysis of phenotypic data collected was performed using R and SAS (SAS Institute Inc., Cary, NC, USA). The phenotypic mean of each RIL across environments was estimated. The proportion of phenotypic variance explained by genetic variation in the NAM for each trait was estimated by fitting terms for the vector of phenotypic data and the matrix of kinship genetic relatedness (estimated from the genomic data) using the heritability package in R (Kruijer *et al.*, 2015). Also, Pearson pairwise correlation between traits was estimated using the residuals derived from fitting a linear model for family and trait phenotypic means;

$$y_{ij} = \mu + \gamma_{\mathrm{i}} + \varepsilon_{\mathrm{ij}} \qquad [1]$$

where y_{ij} is the phenotype, γ_{i} is the term for the NAM families and $\varepsilon_{\mathrm{ij}}$ is the residual term. The BLUP for each phenotype was estimated by fitting RIL, environment, and RIL by environment interaction terms as random using LME4 R package (Bates *et al.*, 2017).

Joint linkage mapping

To characterize the genetic architecture of these traits (STM, LET, PFLW, VLW, HGT, and FLT), joint linkage mapping was performed using trait BLUPs and genomic data using the stepwise regression approach implemented in TASSEL 5.0 (Glaubitz *et al.*, 2014). The approach is based on forward inclusion and backward elimination stepwise methods. The entry and exit limit of the forward and backward stepwise regressions were set at 0.001. Also, the threshold cutoff was set at 1.8 E-6 based on 100 permutations. The genotypic data used for this analysis have been previously described (Bouchet *et al.*, 2017). The JL model was specified as;

$$\mathbf{y} = b_{\mathrm{o}} + \alpha_{\mathrm{f}}\mathrm{u}_{\mathrm{f}} + \sum_{\mathrm{i}=1}^{\mathrm{k}} \mathbf{x}_{\mathbf{i}}\mathrm{b}_{\mathrm{i}} + \mathrm{e}_{\mathrm{i}} \qquad [2]$$

where b_0 is the intercept, u_{f} is the effect of the family of founder *f* obtained in the cross with the common parent (RTx430), α_{f} is the coefficient matrix relating u_{f} to $\mathbf{y}$, b_{i} is the effect of the ith identified locus in the model, $\mathbf{x}_i$ is the incidence vector that relates b_i to $\mathbf{y}$ and k is the number of significant QTL in the final model .

Estimation of QTL effect size and allele frequency

The snpStats package in R (Clayton, 2015) was used to estimate the allele frequency of the QTL. While the proportion of phenotypic variation (PV) explained by each QTL were estimated by fitting a linear model with family and QTL as fixed terms;

$$y_{ijk} = \mu + \gamma_i + \Phi_j + \varepsilon_{ijk} \qquad [3]$$

where y_{ijk} is the phenotype, γ_i is the term for the NAM families, Φ_j is the term for QTL, and ε_{ijk} is the residual term. The sum of squares of QTLs divided by sum of squares total gave the proportion of variance explained by the detected QTL. The additive effect size of the QTL in the population was estimated as the average of the difference between the mean phenotypic values associated with the two-allele class of the QTL. The additive effect size of each QTL was estimated relative to RTx430 (Tian *et al.*, 2011).

Comparison of QTL regions with a priori genes

A list of *a priori* genes was developed for leaf morphology, stem diameter, plant height, and flowering time based on homology with genes underlying these traits in other cereals. This list contained about 146 sorghum homologs in total with 130 of them associated with leaf and stem development, while 16 are known height and flowering time genes in sorghum. A custom R script was used to search for *a priori* genes within 150 kb window around the QTL associations identified in this study.

Results

Phenotypic variation for traits

Significant differences were observed for all traits and the proportion of phenotypic variation explained by genetic variation in the NAM for the traits are 0.56 for HGT, 0.67 for FLT, 0.60 for STM, 0.17 for PFLW, 0.16 for VLW, and 0.50 for LET. Pairwise phenotypic correlations were observed between traits (figure 2.2) with LET having a negative correlation of -0.12 (*P*-value < 0.001) with height, -0.05 (*P*-value < 0.05) with VLW and positive correlation of 0.10 (*P*-value < 0.001) with STM. HGT and

FLT were positively correlated (*r* = 0.13, *P*-value < 0.001), while STM had correlations of 0.23 (*P*-value < 0.001) with PFLW, -0.09 (*P*-value < 0.001) with HGT and 0.10 (*P*-value < 0.001) with LET.

Identified QTL and their effect sizes

The joint linkage analysis identified 13 QTL for STM, 17 QTL for LET, two QTL each for VLW and PFLW, 17 QTL for HGT, and 13 QTL for FLT. About 6 out of the 25 FLT QTL had an additive effect size of 2 days or more (qSbFLT10.78, qSb7.25, qSbFLT3.71, qSbFLT6.57, qSbFLT3.62 and qSbFLT6.79). For HGT, qSbHGT9.570 explained the largest proportion of variation (12%) followed by qSbHGT9.576 (6%). All leaf traits, and stem diameter QTL explained less than 3% of the phenotypic variation. Table 3.3 describes the summary of QTL effect sizes.

Plant and inflorescence morphology genes underlie identified QTL

QTL identified in this study were observed in the proximity of genes known to be associated with inflorescence and plant morphology. For STM, a QTL (qSbSTM7.59) was found about 34-35 kb to the *Dw3* and *YUCCA5* genes. Also, for LET, qSbLET7.63 was found at about 12kb from the sorghum ortholog of rice *OsSPL14* (*Ideal Plant Architecture1*), qSbLET10.60 found at about 1.8 kb from the sorghum ortholog of maize *Thick Tassel Dwarf1* (*CLAVATA1*), qSbLET7.59 was found at about 136 kb from *YUCCA5,* and qSbLET2.63 was found at about 28 kb from a sorghum paralog of maize *Fasciated ear 4* (*Fea4*). Two HGT QTL (qSbHGT7.59 and qSbHGT7.59) were found about 35 kb and 139 kb from *YUCCA5*. Two STM QTL qSbSTM7.51 and qSbSTM3.63 were located at 33 kb and 77 kb from the sorghum paralogs of maize *ROUGH SHEATH2* and *BAD1*. One of LET QTL (qSbLET7.59) was found about 300 kb from *Dw3* and 230 kb from *YUCCA5*. qSbFLT6.79 was found about 99 kb from *Ma6* (Sobic.006G004400), qSbFLT.3_6271 about 30 kb from *SbCN12* (FLOWERING LOCUS T, Sb03g034580), and qSbFLT10.12 about 158 kb from *SbCO*. Likewise, a flowering time QTL qSbFLT6.51 was found at about 107 kb from the sorghum ortholog (Sb06g023770) of the maize leaf morphology gene *yab1/drl1*. The three major sorghum height genes, *Dw1* (Sobic.009G229800), *Dw2* (Sobic.006G067700) and *Dw3* (Sobic.007G163800), were found at about 24 kb, 4.7 kb, and 34 kb from qSbHGT9.57, qSbHGT6.42, and qSbHGT7.59, respectively. QTL qSbHGT7.59 was also found 35 kb from *YUCCA5*, the flavin monooxygenase gene. Also *SbCN4* (Sobic.006G068300) a flowering time gene was found at about 117 kb from qSbHGT6.42 height QTL on chromosome 6.

Discussion

Genetic variation associated with sorghum vegetative traits in the NAM

Genetic variation is essential for breeding. Substantial genetic variation was identified for some of the plant morphology traits evaluated. Four of six traits evaluated in this study (HGT, FLT, STM, and LET) exhibited a high level of heritability genetic variation (Table 3.3). However, the two leaf width traits (PFLW and VLW) showed notably low heritability estimates. Leaf morphology exhibited high heritability in the maize NAM (Cook *et al.*, 2012) and sorghum association panel (Zhao *et al.*, 2016). It is not clear what could have led to the low heritability observed for VLW and PFLW in this study. Phenotyping error is a possible explanation. In this case, automated high-throughput phenotyping on multiple plants would be beneficial to reduce error that may result from manual measurements.

The inverse correlation between PHT and LET showed that a short plant would have more erect leaves. This correlated effect may facilitate adaptation for better light interception under high-density planting. The positive relationship between FLT, and STM, FLT, and HGT is consistent with the expectation that late-flowering plants should have increased stem width and height, and accumulate more biomass (Ashworth *et al.*, 2016).

Few moderate and many small effect size loci associated with sorghum vegetative traits

This study revealed that few loci explaining moderate proportion of phenotypic variation underlie vegetative traits in sorghum. Most of the associated loci, explained less than 5% of the phenotypic variation. Only a single large effect locus was found to be associated with HGT (qSbHGT9.57, *Dw1* QTL explaining 12% of PV). All QTL associated with leaf traits were of small effect explaining less than 3% of phenotypic variation. In the maize NAM population, small effect loci were found to be associated with both leaf erectness and width (Tian *et al.*, 2011). Earlier QTL mapping for leaf erectness in sorghum using biparental population identified the average of four QTL per population with most QTL having an effect size of about 15% (Truong *et al.*, 2015). However, the estimated effect size of these loci could have been inflated due to the Beavis effect (small population) (Xu, 2003). The estimated effects in this study are expected to be more accurate due to the large number of RILs in the NAM population (2200 RILs). A positive relationship was observed between heritability and the number of associated QTL since only the two low heritable leaf width traits were the ones with the least QTL number. This showed that the presence of high genetic variation is essential for QTL mapping.

Vegetative traits were associated with genes underlying plant and inflorescence development

QTL identified in this study were found to be in proximity of genes controlling plant morphology and inflorescence morphology. One important QTL (qSbHGT7.59) identified in this study was found to be pleiotropic with HGT, STM, and LET. The QTL was found at about 34 kb from *Dw3* and *YUCCA5* genes. Though there is limited knowledge about the genetic basis of leaf erectness in sorghum, associations underlying it have been found in the qSbHGT7.59 region (Hart *et al.*, 2001; Truong *et al.*, 2015; Zhao *et al.*, 2016). *Dw3* is an auxin transporter gene whose mutation led to dwarf plants in sorghum (Multani *et al.*, 2003). The pleiotropic effect of *Dw3* on leaf erectness was recently described (Truong *et al.*, 2015). Dwarf lines carrying *dw3* duplication to be more erect, while tall *Dw3* revertants (those that changed from dwarf to tall stature) were less erect. *YUCCA5* is a flavin monooxygenase gene involved auxin biosynthesis (Dai *et al.* 2013). *YUCCA5* homologs have been found to control inflorescence morphology in maize (Gallavotti *et al.*, 2008) and rice (Yamamoto *et al.*, 2007), and have been proposed as candidates for inflorescence morphology in sorghum (Brown *et al.*, 2008). In addition, a moderate effect loci underlying LET, *OsSPL14* (*Ideal Plant Architecture1, IPA1*) was found about 12 kb from a LET QTL (qSbLET7.63). *IPA1* has been found to play a critical role in both plant architecture and inflorescence morphology in rice (Si *et al.*, 2016). Similarly, prior to this study, little is also known about the underlying genetic basis of stem diameter in sorghum (Mantilla Perez *et al.*, 2014). But sorghum paralogs of maize *ROUGH SHEATH2* and *BAD1* were found to underlie some of the STM QTL identified in this study.

Conclusion

This study provided insights into the genetic basis of plant architecture in sorghum. Leaf erectness and stem diameter were under the control of common moderate effect loci. A pleiotropic QTL (qSbHGT7.59) that co-localizes with YUCCA5 and Dw3 regions was found to be associated with HGT, STM, and LET. The colocalization and/or pleiotropy of these QTL and correlation relationships between the three traits showed that selection for dwarf height in sorghum might have led to an indirect selection for increased leaf erectness and increased stem diameter. These QTL can be used to develop molecular markers to facilitate simultaneous selection and improvement of vegetative morphology. The limited number of QTL and low heritability estimates found to be associated with leaf width traits in this study calls for a thorough investigation of the traits with more accurate high-throughput phenotyping methods. Interestingly, some of the QTL underlying vegetative traits in this study were found near candidate genes for inflorescence morphology. This suggest possible pleiotropy of these genes in both vegetative and inflorescence development biology for example liguless gene in rice *OsLG1* (Ishii *et al.*, 2013). QTL identified in this study can help facilitate marker development for ideotypic breeding and further the study of the underlying genes.

Tables and Figures

Table 3.1: Sorghum nested association mapping population founders, their countries of origin and number of recombinant inbred lines (RILs) present in each family.

Founder	Origin	Founder Type	RILs
RTX430	Texas A & M University	Common Parent	-
P898012	Purdue University	Diverse Founder	213
Ajabsido	Sudan	Diverse Founder	214
Macia	ICRISAT	Diverse Founder	231
SC1103	Nigeria	Diverse Founder	231
SC1345	Mali	Diverse Founder	231
SC265	Burkina Faso	Diverse Founder	232
SC283	Tanzania	Diverse Founder	223
SC35	Ethiopia	Diverse Founder	208
SC971	Puerto Rico, United States	Diverse Founder	233
Segaolane	Botswana	Diverse Founder	204

Table 3.2: Multi-environmental phenotypic evaluation of sorghum nested association-mapping population. Phenotypic traits evaluated are Leaf erectness (LET), pre-flag leaf width (PFLW), vegetative leaf width (VLW), stem diameter (STM), height (HGT), and flowering time (FLT).

Location	**Climate**	**Year**	**Precipitation (mm)** **Oct – Oct***	**Environment Code**	**Traits measured**
Manhattan, KS	Humid Continental	2014	698	MN14	LET, PFLW, VLW, STM, FLT
Hays, KS	Semi-Arid	2014 (Upland)	639	HA14	STM, HGT
Manhattan, KS	Humid Continental	2015	998	MN15	LET, VLW, FLT
Hays, KS	Semi-Arid	2015 (Bottomland)	513	HI15	LET, VLW, PFLW, HGT
Hays, KS	Semi-Arid	2015 (Upland)	513	HD15	VLW

* National Oceanic and Atmospheric Administration, U.S. Department of Commerce.

Table 3.3: Narrow sense heritability, additive effect size (AES) and proportion of phenotypic variation explained by QTL.

Trait	**h^2**	**Range of PPVE (%)**	**Range of AES**
HGT	0.54	0.2 to 12	-12 to 7 (cm)
FLT	0.71	0.1 to 6	-1 to 4 (days)
STM	0.60	0.3 to 1	-1 to 3 (mm)
LET	0.50	0.3 to 3	-7 to 4 (degree)
PFLW	0.17	0.8 to 1.8	-5 to -3 (mm)
VLW	0.16	0.6 to 0.9	-1 to 13 (mm)

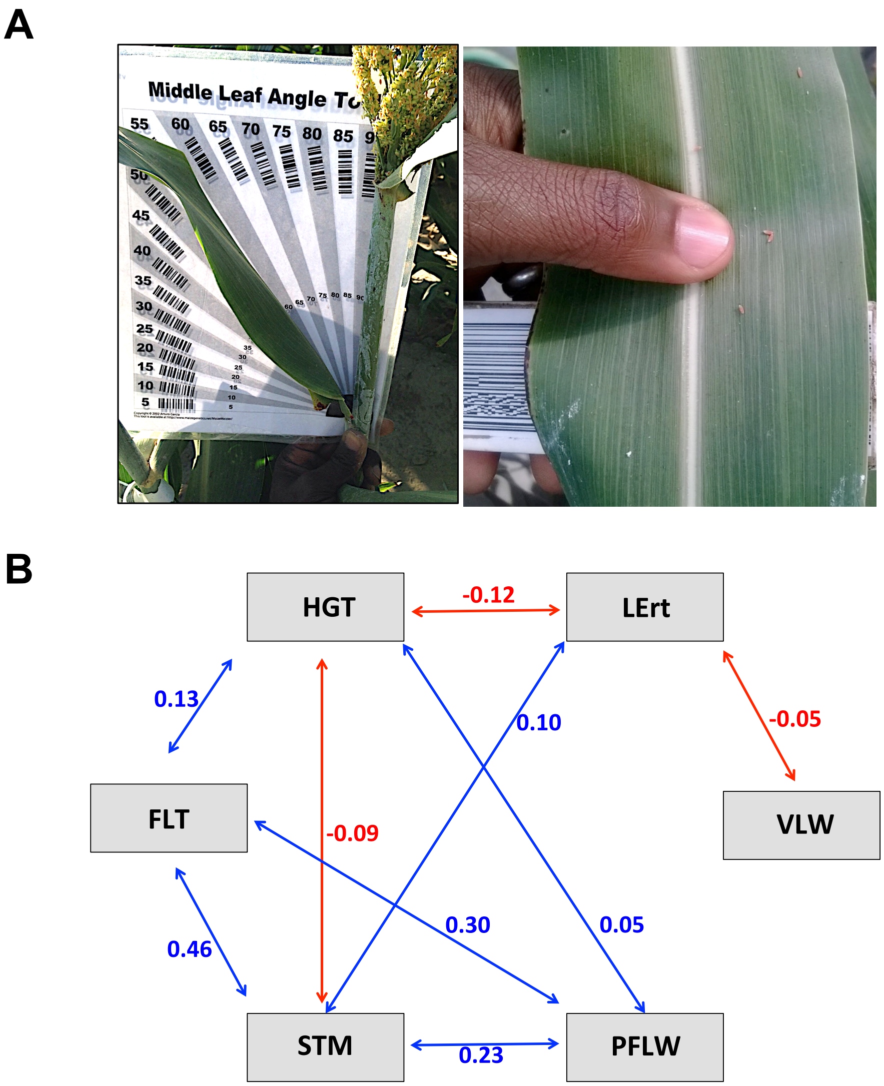

Figure 3.1: Phenotypic variation in vegetative morphology.

(A) Leaf erectness and leaf width measurements. A barcode protractor and barcode ruler were both used for the measurements of these traits to ensure accurate and fast data acquisition. (B) Pairwise correlations among traits after accounting for family effect (Correlation of residuals of a linear model with a fixed family term). The correlation values shown are significant at *P*-value < 0.01. Blue lines indicate positive relationships while red lines indicate negative relationships.

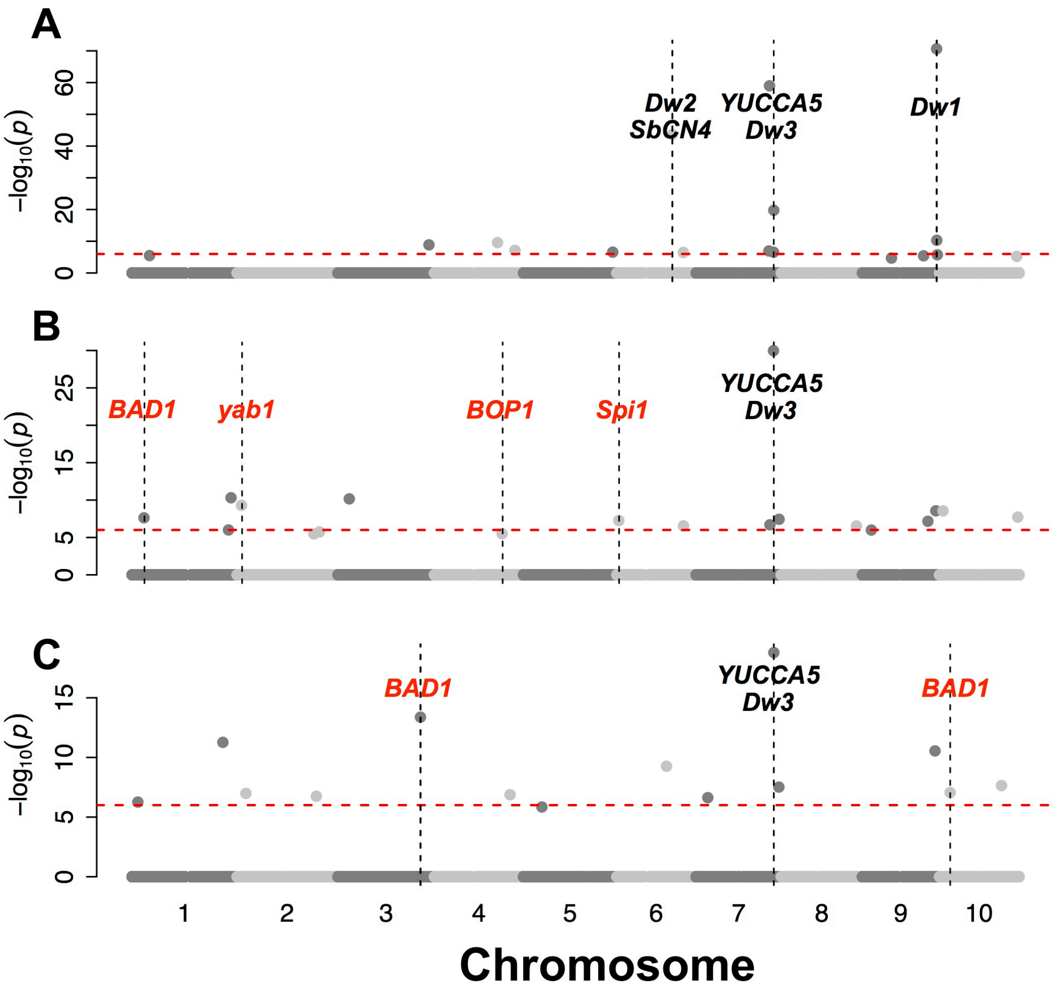

Figure 3.2: QTL mapping for vegetative morphology using joint linkage model.

Genomic location of associations with (A) plant height, (B) leaf erectness, and (C) stem diameter. The dashed red lines are the Bonferroni significance threshold (P-value < 0.05) estimated from 100 permutations. *A priori* candidate genes that colocalize with QTL within 150 kb are noted as follows. Black text indicates putative sorghum orthologs of *a priori* candidate genes while red text indicates paralogs.

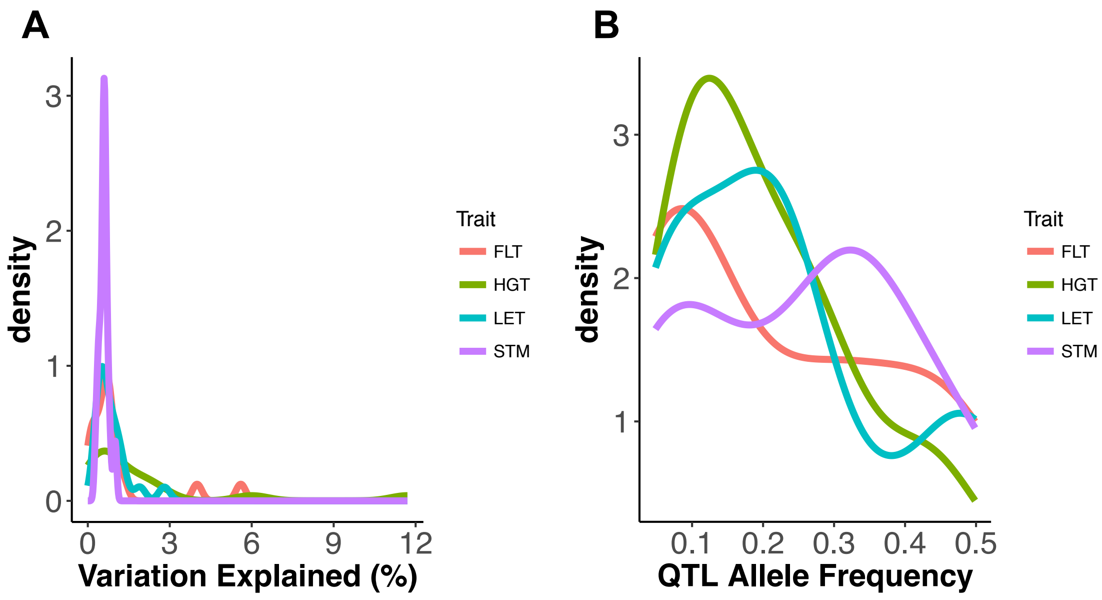

Figure 3.3: Distribution of QTL effects and allele frequencies.

(A) Density plots summarizing the distribution of QTL effect sizes for each trait. Percent variance explained of the QTL was estimated from linear models with a fixed main effect of family and fixed main effect of QTL. (B) Density plots summarizing the distribution of QTL allele frequencies for each trait. FLT: Flowering time, HGT: Plant height, LET: Leaf erectness, STM: Stem diameter.

References

Ashworth MB, Walsh MJ, Flower KC, Vila-Aiub MM, Powles SB. **2016**. Directional selection for flowering time leads to adaptive evolution in *Raphanus raphanistrum* (Wild radish). *Evolutionary Applications* **9**: 619–629.

Bates D, Maechler M, Bolker B, Walker S, Christensen RHB, Singmann H, Dai B, Eigen C. **2017**. Fitting linear mixed-effects models using lme4. *Journal of statistical software* **67**: 1–113.

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.

Boyko AR, Quignon P, Li L, Schoenebeck JJ, Degenhardt JD, Lohmueller KE, Zhao K, Brisbin A, Parker HG, Vonholdt BM, *et al.* 2010. A Simple Genetic Architecture Underlies Morphological Variation in Dogs. *PLoS Biology* **8**: 1–13.

Brown PJ, Rooney WL, Franks C, Kresovich S. **2008**. Efficient Mapping of Plant Height Quantitative Trait Loci in a Sorghum Association Population With Introgressed Dwarfing Genes. *Genetics* **180**: 629–637.

Brown PJ, Upadyayula N, Mahone GS, Tian F, Bradbury PJ, Myles S, Holland JB, Flint-Garcia S, McMullen MD, Buckler ES, *et al.* 2011. Distinct genetic architectures for male and female inflorescence traits of maize. *PLoS Genetics* **7**: e1002383.

Carlborg R, Jacobsson L, Hgren PÅ, Siegel P, Andersson L. **2006**. Epistasis and the release of genetic variation during long-term selection. *Nature Genetics* **38**: 418–420.

Clayton D. **2015**. snpStats: SnpMatrix and XSnpMatrix classes and methods. R package version 1.28.0.

Cook JP, McMullen MD, Holland JB, Tian F, Bradbury P, Ross-Ibarra J, Buckler ES, Flint-Garcia S a. 2012. Genetic Architecture of Maize Kernel Composition in the Nested Association Mapping and Inbred Association Panels. *Plant Physiology* **158**: 824–834.

Cresko WA, Amores A, Wilson C, Murphy J, Currey M, Phillips P, Bell MA, Kimmel CB, Postlethwait JH. **2004**. Parallel genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback populations. *Proceedings of the National Academy of Sciences of the United States of America* **101**: 6050–5.

Doebley J. **2004**. The Genetics of Maize Evolution. *Annual Review of Genetics* **38**: 37–59.

Donald CM. **1968**. The breeding of crop ideotypes. *Euphytica* **17**: 385–403.

Duvick DN. **2005**. Genetic progress in yield of United States maize (Zea mays L.). *Maydica* **50**: 193–202.

Gallavotti A, Barazesh S, Malcomber S, Hall D, Jackson D, Schmidt RJ, McSteen P. **2008**. sparse inflorescence1 encodes a monocot-specific YUCCA-like gene required for vegetative and reproductive development in maize. *Proceedings of the National Academy of Sciences of the United States of America* **105**: 15196–15201.

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 (NA Tinker, Ed.). *PLoS ONE* **9**: e90346.

Hammer GL, Dong Z, McLean G, Doherty A, Messina C, Schussler J, Zinselmeier C, Paszkiewicz S, Cooper M. 2009. Can changes in canopy and/or root system architecture explain historical maize yield trends in the U.S. corn belt? *Crop Science* **49**: 299–312.

Hart GE, Schertz KF, Peng Y, Syed NH. 2001. Genetic mapping of Sorghum bicolor (L.) Moench QTLs that control variation in tillering and other morphological characters. *Theoretical and Applied Genetics* **103**: 1232–1242.

Ishii T, Numaguchi K, Miura K, Yoshida K, Thanh PT, Htun TM, Yamasaki M, Komeda N, Matsumoto T, Terauchi R, *et al.* 2013. OsLG1 regulates a closed panicle trait in domesticated rice. *Nature genetics* **45**: 462–5, 465–2.

Kashiwagi T, Togawa E, Hirotsu N, Ishimaru K. 2008. Improvement of lodging resistance with QTLs for stem diameter in rice (Oryza sativa L.). *Theoretical and Applied Genetics* **117**: 749–757.

Khush GS. 2001. TIMELINE: Green revolution: the way forward. *Nature Reviews Genetics* **2**: 815–822.

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**.

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

Kruijer W, Boer MP, Malosetti M, Flood PJ, Engel B, Kooke R, Keurentjes JJB, Van Eeuwijk FA. 2015. Marker-based estimation of heritability in immortal populations. *Genetics* **199**: 379–398.

Li C, Li Y, Shi Y, Song Y, Zhang D, Buckler ES, Zhang Z, Wang T, Li Y. 2015. Genetic Control of the Leaf Angle and Leaf Orientation Value as Revealed by Ultra-High Density Maps in Three Connected Maize Populations (R Wu, Ed.). *PLOS ONE* **10**: e0121624.

Mantilla Perez MB, Zhao J, Yin Y, Hu J, Salas Fernandez MG. 2014. Association mapping of brassinosteroid candidate genes and plant architecture in a diverse panel of Sorghum bicolor. *Theoretical and Applied Genetics* **127**: 2645–2662.

McCouch S, Katayose Y, Ashikari M, Yamanouchi U, Monna L. 2004. Diversifying Selection in Plant Breeding. *PLoS Biology* **2**: e347.

Morrell PL, Buckler ES, Ross-Ibarra J. 2011. Crop genomics: advances and applications. *Nature Reviews Genetics*.

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 (New York, N.Y.)* **302**: 81–4.

Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, Costich DE, Buckler ES. **2009**. Association Mapping: Critical Considerations Shift from Genotyping to Experimental Design. *The Plant Cell* **21**: 2194–2202.

Paterson AH. **2013**. Genomics of the Saccharinae. *Genomics of the Saccharinae*: 1–567.

Qi J, Qian Q, Bu Q, Li S, Chen Q, Sun J, Liang W, Zhou Y, Chu C, Li X, *et al.* 2008. Mutation of the Rice Narrow leaf1 Gene, Which Encodes a Novel Protein, Affects Vein Patterning and Polar Auxin Transport 1[OA]. *Plant physiology* **147**: 1947–1959.

Si L, Chen J, Huang X, Gong H, Luo J, Hou Q, Zhou T, Lu T, Zhu J, Shangguan Y, *et al.* 2016. OsSPL13 controls grain size in cultivated rice. *Nature Genetics* **48**: 447–456.

Strable J, Wallace JG, Unger-Wallace E, Briggs S, Bradbury P, Buckler ES, Vollbrecht E. **2017**. Maize YABBY Genes drooping leaf1 and drooping leaf2 Regulate Plant Architecture. *The Plant Cell*: tpc.00477.2016.

Tian F, Bradbury PJ, Brown PJ, Hung H, Sun Q, Flint-Garcia S, Rocheford TR, McMullen MD, Holland JB, Buckler ES. **2011**. Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nature genetics* **43**: 159–62.

Truong SK, McCormick RF, Rooney WL, Mullet JE. **2015**. Harnessing Genetic Variation in Leaf Angle to Increase Productivity of Sorghum bicolor. *Genetics* **201**: 1229–38.

Wang Y, Li J. **2008**. Molecular Basis of Plant Architecture. *Annu. Rev. Plant Biol* **59**: 253–79.

Xu S. **2003**. Theoretical Basis of the Beavis Effect. *Genetics* **165**: 2259–2268.

Yamamoto Y, Kamiya N, Morinaka Y, Matsuoka M, Sazuka T. **2007**. Auxin biosynthesis by the YUCCA genes in rice. *Plant physiology* **143**: 1362–1371.

Zhao J, Mantilla Perez MB, Hu J, Salas Fernandez MG. **2016**. Genome-Wide Association Study for Nine Plant Architecture Traits in Sorghum. *The Plant Genome* **9**: 0.

Zhao SQ, Xiang JJ, Xue HW. **2013**. Studies on the rice leaf inclination1 (LC1), an IAA-amido synthetase, reveal the effects of auxin in leaf inclination control. *Molecular Plant* **6**: 174–187.

Chapter 4 - Population and Quantitative Genomic Analysis of Clinal Adaptation in Nigerian Sorghum

Abstract

Sorghum landraces have adapted to different environments, providing genetic diversity useful for crop improvement. Nigeria harbors abundant sorghum diversity, however, inadequate understanding of genetic diversity, population structure, and adaptive loci limits germplasm utilization. In this study, 607 Nigerian sorghum accessions were characterized at > 400,000 SNPs and compared regional West African germplasm and a global reference germplasm. Nigerian germplasm has a substantial level of genetic diversity (π) compared to global (~ 98%) and West African (~ 92%) germplasm. Discriminant analysis of principal components identified three distinct genetic groups that are moderately genetically differentiated from each other. Linkage disequilibrium in the Nigerian germplasm was slightly slower than that of the global germplasm, which indicated possible lower mapping resolution in the Nigerian germplasm. A genome scan for signatures of adaptation and genome-wide association studies identified signals in the proximity of candidate genes underlying flowering time, height, and inflorescence architecture. Results suggest that the genomic diversity in Nigerian landraces was shaped by clinal adaptation across a climatic gradient and can provide genetic resources for crop improvement.

Introduction

Intraspecific phenotypic variation exists for adaptive traits across a climatic gradient in natural populations. Latitudinal variations have been observed for seed dormancy, cold tolerance, height, and flowering time in Arabidopsis (Kronholm *et al.*, 2012; Samis *et al.*, 2012; Debieu *et al.*, 2013). Likewise, in cultivated crop species, geographical distribution from tropical to cold temperate conditions became possible through adaptation of flowering time to local conditions (Camus-Kulandaivelu *et al.*, 2006; Ducrocq *et al.*, 2008; Buckler *et al.*, 2009). Furthermore, local adaptation of traditional varieties has played essential roles in ensuring marginal yield under adverse climatic conditions in smallholder farmers' fields and low input agricultural systems (Mercer *et al.*, 2012; Feitosa Vasconcelos *et al.*, 2013). These locally adapted varieties possess alleles that can be beneficial for the development of better-adapted lines in crop improvement to ensure food security (Zeven, 1998; Soler *et al.*, 2013; Lasky *et al.*, 2015). Therefore, we need to improve our understanding of adaptive genetic diversity in crop species for efficient utilization in crop improvement.

Characterization of genetic diversity and population structure of natural crop species populations is important for (Djè *et al.*, 2000; Manzelli *et al.*, 2007; Samis *et al.*, 2012; Soler *et al.*, 2013; Yoder *et al.*, 2014). The genetics of local and clinal adaptation has been widely studied using population genomic approaches (Umina *et al.*, 2005; Zhen & Ungerer, 2008; Samis *et al.*, 2012; Yoder *et al.*, 2014). Some of the approaches involve genome scans for fingerprints of selection at linked neutral loci (Siol *et al.*, 2010). Another widely used approach is genome-wide associations between genomic variants and climatic factors (Hancock *et al.*, 2011; Yoder *et al.*, 2014). Population genomic tools have helped improve our understanding of phenotypic evolution in crop species like maize (Ducrocq *et al.*, 2008; Hufford *et al.*, 2012; van Heerwaarden *et al.*, 2012), rice (Olsen *et al.*, 2006), and sorghum (Morris *et al.*, 2013; Lasky *et al.*, 2015; Zhang *et al.*, 2015)

Sorghum bicolor is an essential staple cereal crop in dryland regions of the world and it has adapted to a wide range of climatic environments with intraspecific phenotypic variation across clines for flowering time, plant architecture and inflorescence architecture (Thurber *et al.*, 2013; Morris *et al.*, 2013; Lasky *et al.*, 2015; Zhang *et al.*, 2015). For instance, in West Africa there is a strong north-south climatic gradient, from semiarid grasslands bordering the Sahara desert in the north (Sahelian zone), through subhumid savannah (Sudanian zone), to humid rainforest in the south (Guinean zone). Sorghum phenotypic variation varies along this climatic gradient. Open panicle sorghum types are predominant in the humid regions, while semi-compact to compact panicle types are predominant in the semi-arid and arid regions (Harlan, 1992). The diversity of climatic zones often varies from country to country in the region. The Nigerian geographical landscape is divided into about eight agroclimatic zones based on precipitation patterns (Oyenuga, 1967; Sowunmi & Akintola, 2010). Sorghum is a major cereal in the northern (Sudano-Sahelian) regions of Nigeria, which are characterized by prolonged dry seasons and hot weather. Nigeria is the second largest sorghum producer globally, with 5-10 million tons (Mg) of production per year (WSP, 2017).

The Nigerian population stands at about 190 million and is expected to be larger than the US population by 2050 (CNN, 2017), which raises concerns for food security. In addition, increasing temperature and erratic rainfall due to climate change also threaten agricultural food production. Therefore, it is important to develop better-adapted and high yielding sorghum varieties for farmers. However, the genetic diversity of sorghum in Nigeria is still not characterized compared to other West African germplasm (Ezeaku *et al.*, 1999; Ezeaku & Gupta, 2004; Deu *et al.*, 2008; Tesso *et al.*, 2008; Bezançon *et al.*, 2009; Vigouroux *et al.*, 2011). Characterizing Nigerian sorghum germplasm diversity can facilitate the identification of potential sources of adaptive traits and genetic diversity relevant to crop improvement. The objectives of this study were thus: (1) to evaluate the genetic diversity of the Nigerian germplasm in relation to West African and global sorghum, (2) to examine whether traits have been shaped by local adaptation in response to climatic factors in Nigeria, and (3) to identify the genomic regions responsible for local adaptation. To understand the genetic diversity of Nigerian germplasm, 607 Nigerian accessions were sequenced using GBS. Sequence data were combined with previous sequencing data from 1785 georeferenced global accessions (Lasky *et al.*, 2015) to perform population and quantitative genomics analysis.

Materials and Methods

Plant materials

A set of 553 Nigerian accessions was obtained from the USDA National Plant Germplasm System (NPGS) (https://www.ars-grin.gov/). From another set of 1943 georeferenced global accessions previously sequenced (Lasky *et al.*, 2015), sequence information from 158 Nigerian accessions was obtained and combined with the Nigerian NPGS set. Duplicated accessions and those with the US sorghum conversion (SC) program identity number in the NPGS database were removed from the Nigerian germplasm. Thus, only 607 Nigerian accessions (of which only 443 were georeferenced) and 1785 georeferenced global accessions were used for downstream analysis. Precipitation maps were generated using Nigerian average annual precipitation data (from 1960 to 1990) obtained from WorldClim 1.4 with Raster package in R (Hijmans 2016). The distribution of georeferenced Nigerian accessions across precipitation zones was plotted using the raster package in R (Hijmans, 2016). The 553 NPGS accessions were raised in the greenhouse for two weeks and about 50 mg of fresh leaf tissue was collected from each accession into 96 well plates. A control well was left empty on each plate. Leaf tissue was lyophilized (Labconco Freeze Dryer, Kansas City, MO USA) for two days and then ground using 96-well plate plant tissue grinder (Retsch Mixer Mill, Haan, Germany). Genomic DNA was extracted using BioSprint 96 DNA Plant Kit (QIAGEN, Valencia CA, USA), quantified using Quant-iTTM PicoGreen® dsDNA Assay Kit (ThermoFisher Scientific, Waltham MA, USA) followed by normalization to 10ng/ul.

Genotyping-By-Sequencing Pipeline

The GBS approach described by (Elshire *et al.*, 2011) was used for GBS of 553 Nigerian accession. Individual DNA samples were digested using ApeKI restriction enzyme (NEB R0643L) followed by ligation of barcode and common adapters ligation using T4 DNA ligase (NEB M0202L). The ligated libraries were pooled (96-plex libraries) and purified with the QIAquick PCR purification kit (QIAGEN, Valencia CA, USA). Library size distribution was obtained using a Bioanalyzer (Agilent Technologies 2100, Santa Clara CA, USA). The 384-plex library was obtained by pooling four 96-plex libraries. Libraries were sequenced using single end 100-cycle sequencing using Illumina

HiSeq2500 (Illumina, San Diego CA, USA) at the University of Kansas Medical Center, Kansas City MO, USA. Raw reads for Nigerian germplasm were combined with raw reads obtained for 1943 accessions (Lasky *et al.*, 2015) and tags were aligned to the sorghum reference genome v3.0 obtained from (Goodstein *et al.*, 2012) using Burrow Wheeler Alignment algorithm (Li & Durbin, 2009). SNP calling was performed using TASSEL 5.0 (Glaubitz *et al.*, 2014). Monomorphic markers and singletons were removed prior to the imputation of missing data using BEAGLE 4.0 (Browning & Browning, 2013).

Linkage disequilibrium, neighbor-joining, and principal component analysis

Linkage disequilibrium decay for the genomic data for Nigerian and global germplasm was estimated by PopLDdecay (BGI-shenzhen, 2017), with minor allele frequency parameter set at 0.05, and smoothing was done by the spline function in R. Phylogenetic analysis for neighbor-joining tree was performed using TASSEL 5.0 and APE (Analyses of Phylogenetics and Evolution) package in R (Paradis *et al.*, 2004). Only two accessions from the global data did not have country information. The phylogenetic tree was constructed using 311,786 SNPs from the West African data after filtering for monomorphic and singleton markers. Likewise, monomorphic and singletons were removed from Nigerian germplasm remaining 268,326 SNPs. For the discriminant analysis of principal components (DAPC) analysis, the find clusters function in Adegenet package in R (Jombart *et al.*, 2010) was first used to infer the possible number of groups or clusters in the 607 Nigerian accessions.

Genetic diversity and population differentiation

The estimate of nucleotide divergence was used as a measure of genetic diversity using --site-pi and –window-pi options in VCF tools (Danecek *et al.*, 2011). This was performed for the Nigerian germplasm, West African germplasm (made up of Benin, Togo, Ghana, Senegal, Gambia, Burkina Faso, Sierra Leone, Niger, and Cameroon a border country to Nigeria on the east). Furthermore, the extent of population differentiation between Nigeria and the global germplasm and West African germplasm was estimated using --weir-fst-pop parameter (Weir and Cockerham's F_{ST}) in VCF tools (Danecek *et al.*, 2011).

Genome scans for signature of adaptation

Genomic signatures of adaptation in the Nigerian germplasm were identified using PCAdapt R package (Luu *et al.*, 2017), with Nigerian accessions (601) and global reference accessions (1941) analyzed separately. An optimal cluster number of 11 and 17 was identified through the scree plot for the Nigerian and global reference sets, respectively. Outliers SNPs (putatively under selection) were identified using the PCAdapt function (Luu *et al.*, 2017) with parameters for optimal cluster groups and minor allele frequency of 0.01 for both Nigerian and global reference data sets. An *a priori* candidate gene list of sorghum orthologs was compiled based on the literature search for genes underlying flowering time, inflorescence, and plant architecture in cereals.

Genome-wide association mapping for environmental and phenotypic data

Climate data (annual mean temperature, mean temperature wettest quarter, mean temperature driest quarter, mean temperature warmest quarter, mean temperature coldest quarter, annual precipitation, precipitation wettest quarter, and precipitation driest quarter from 1960 to 1990) were obtained from WorldClim 1.4 (Worldclim.org) using the Raster package in R (Hijmans, 2016)based on the coordinate (latitude and longitude) information for each of the 438 georeferenced Nigerian accessions. Passport data for flowering time, plant height, and panicle length for the Nigerian accessions were obtained from the USDA National Plant Germplasm System (NPGS) (https://www.ars-grin.gov/) database. Correlations between three adaptive traits (flowering time, panicle length, and plant height) and climatic factors (temperature and precipitation) were estimated. Genome-wide association mapping (GWAS) was performed using the climate data as the phenotypic data. Mixed linear model (GAPIT MLM; (Lipka *et al.*, 2012)) and multi-locus mixed linear model (MLMM; (Segura *et al.*, 2012)) that both accounted for population structure (Q, fixed effects) and a random polygenic term (K, representing kinship relationship matrix) were used to perform GWAS. The MLMM approach performs stepwise regression involving both forward and backward regressions. A total 189,750 were used in the GWAS analysis and coded as 2 and 0 for homozygous SNPs and 1 for heterozygous SNPs. Bonferroni correction of 2.6e-07 (α/total number of markers; where α = 0.05) was used to determine the cut-off threshold for the associations.

Results

Nigerian and global germplasm

A total of 431,698 SNPs were obtained for the total set of 2542 worldwide accessions. In the Nigerian genomic data, after removing monomorphic markers, singletons, and doubletons 189,750 SNPs were retained. This corresponds to an average of 1 SNP per 4 kb. In addition, a West African subset (325 accessions) of the genomic data was also created having about 311,786 SNPs after removing monomorphic markers and singletons. SNP density was found to be high around sub-telomeric regions, while reduced in sub-centromeric regions (Figure 4.2). About 51% of the Nigerian genomic data is composed of SNPs with minor allele frequencies (MAF) < 0.01. By contrast, 46% of the West African SNPs have MAF < 0.01, 36% of the global reference SNPs have MAF < 0.01, and 37% of the whole population SNPs have MAF < 0.01% (Figure 4.3A). Inbreeding coefficients of global (0.83) and west African accessions (0.82) were higher than that of Nigerian accessions (0.80) (P-value < 0.001 and P-value < 0.01, respectively). LD decayed to half its initial value at 12 kb and to background level ($r^2 < 0.1$) at 180 kb in the Nigerian germplasm (Figure 4.3B). The West African germplasm had the slowest LD decay rate compared to the Nigerian and global germplasm (Figure 4.3B).

Germplasm genetic diversity and relatedness

The average nucleotide diversity across 1 kb windows for global germplasm (without Nigeria), West African germplasm, and Nigerian germplasm are 0.00046, 0.00049 and 0.00045 respectively. The Nigerian germplasm had a negative Tajima's *D* value of -0.2 while the global and West African accessions had Tajima's *D* values of 0.1 and 0.2, respectively (using 1 kb windows). Neighbor-joining analysis showed that Nigerian accessions grouped together with West African accessions in the global germplasm (Figure 4.4A). Clustering by botanical race was also observed in the Nigerian germplasm neighbor-joining tree (Figure 4.4C). DAPC analysis identified 3 distinct genetic groups across geographical zones (Figure 4.5A–C). The DAPC groups were genetically differentiated (F_{ST}) from each other as follows: group 1 versus group 2 (F_{ST} of 0.21), group 1 versus group 3 (F_{ST} of 0.18), group, and 2 versus group 3 (F_{ST} of 0.22).

Relationships between adaptive traits and climatic factors

Significant correlations were observed for flowering time with an annual temperature (-0.14, *P*-value < 0.01) (Figure 4.6), mean temperature wettest quarter (-0.15, *P*-value <0.01), mean temperature warmest quarter (-0.15, *P*-value < 0.01), mean temperature coldest quarter (-0.11, *P*-value < 0.01). Plant height had correlation values of 0.28 (*P*-value < 0.001), 0.24 (*P*-value < 0.001), 0.29 (*P*-value < 0.001), 0.18 (*P*-value < 0.01), and 0.20 (*P*-value < 0.01) with a mean temperature driest quarter, mean temperature coldest quarter, annual precipitation (Figure 4.6), precipitation in the wettest quarter and precipitation in the driest quarter respectively. Panicle length had correlations of 0.12 (*P*-value < 0.05), 0.24 (*P*-value < 0.001), 0.24 (*P*-value < 0.001) with mean temperature coldest quarter, annual precipitation (Figure 4.6), and precipitation in the warmest quarter respectively. Latitude had significant negative relationships with plant height and panicle length (-0.33 and -0.21 *P*-value < 0.001) respectively (Figure 4.7). However, there was no relationship between latitude and flowering time. Redundancy analysis showed that climatic factors and space (latitude and longitude) both explained only 5% of the genetic variation in the Nigerian germplasm. However, when phenotypic information with no missing data (209 accessions) was later included in the model, climatic factors, space, and phenotypes explained 9% of the SNP variation (Figure 4.8).

Genome scans for selection, GWAS, and allele distributions

Two major selective sweeps were observed on chromosome 6 as well as other outliers distributed across the genome (Figure 4.9A). For the GWAS result, significant association signals were found for panicle length on chromosome 3 (Figure 4.9B). Association signal was found on chromosome 3 for plant height and chromosome 2 for panicle length (figure 4.9C). For flowering time, significant associations were found on chromosomes 3, 4, 6, 8, and 10 (Figure 4.9D). Among the climatic variables analyzed by GWAS, significant associations were only observed for annual precipitation and mean precipitation in the driest quarter (Figures 4.9E–F). Alleles of NAM inflorescence QTL had a distinct agroclimatic pattern of distribution (Figures 4.10A–B).

Discussion

Genetic diversity in Nigerian sorghum germplasm

The evaluation and comparison of the genetic diversity of Nigerian sorghum germplasm with West African and global germplasm in this study provided an informative description of the potentials of the Nigerian germplasm as a source of plant genetic resources for crop improvement. The extent of genetic diversity inherent in the Nigerian germplasm ($\pi = 4.5 \times 10^{-4}$ per 1 kb windows) was found to be slightly lower than that of the global germplasm ($\pi = 4.6 \times 10^{-4}$ [per 1 kb windows], P-value < 0.001) and West Africa ($\pi = 4.9 \times 10^{-4}$ [per 1 kb windows] P-value < 0.001). This level of genetic diversity observed in the Nigerian germplasm can be considered substantial relative to that of the global germplasm that has a larger number of accessions. A similar estimate of genetic diversity was found in the Nigerien germplasm (country: Niger, $\pi = 4.6 \times 10^{-4}$ (per 1 kb windows), (Maina *et al.* 2017, in review)). The level of genetic differentiation between the Nigerian germplasm and West African germplasm ($F_{ST} = 0.007$) is 10 times lower than the level of genetic differentiation between the Nigerian germplasm and global germplasm ($F_{ST} = 0.07$). Neighbor-joining tree analysis also showed that majority of the Nigerian accessions clustered together with accessions from West Africa (Figure 2B) while the Nigerian germplasm showed distinct clusters based on botanical races (Figure 2C). The genetic relatedness of the Nigerian germplasm to the rest of the West African germplasm can facilitate efficient exchange of genetic materials between Nigerian breeding programs and other West African national breeding programs.

Discriminant analysis of principal components grouped the Nigerian accessions into three major genetic groups (figure 4) with group 2 (mostly represented by Guinea and Guinea derivatives) found to be more prevalent in the middle-belt and humid southern part of Nigeria while groups 1 and 3 (Caudatum, Durra, and their derivatives) were found to be more prevalent in the northeastern and northwestern parts of Nigeria. These groups were found to be moderately genetically distant from each other (F_{ST} 0.18–0.22), which might suggest an agroclimatic and a racial pattern of distribution and differentiation within the Nigerian germplasm. This is consistent with prior studies where sorghum germplasm was structured according to botanical race and geography (de Oliveira *et al.*, 1996; Barnaud *et al.*, 2007).

Differences in linkage disequilibrium decay were observed between the West African and Nigerian germplasm with the faster LD decay in the Nigerian germplasm. This difference may be attributed the effect of population structure which is higher in the West African germplasm. Genome-wide, LD decay rate in Nigerian germplasm was slower than the previous findings using global germplasm ((Hamblin *et al.*, 2005)[r^2 below 0.1 at 15–20 kb]; (Bouchet *et al.*, 2012) [$r^2 = 0.18$ (within 0–10 kb interval) and $r^2 = 0.03$ (within 100 kb – 1 Mb interval)]; (Morris *et al.*, 2013) [half of its initial value by 1 kb and to background levels ($r^2 < 0.1$) within 150 kb]), which were on more diverse panels. The LD decay rate, coupled with the rich genetic diversity present in the Nigerian germplasm, makes it a potential resource for trait mapping.

Sorghum phenotypic variation in Nigeria has been shaped by climatic factors

Intraspecific variations are often associated with phenotypes conferring adaptation across agroclimatic regions. In this study, significant relationships were found between adaptive traits and climatic factors and latitude. The negative relationship between plant height and panicle length and positive relationship of plant height with annual precipitation (0.29 and 0.24, P-values < 0.001) indicate that sorghum plants originating from lower latitudes in Nigeria are taller and have longer panicles than sorghum plants originating from higher latitudes. This is plausible since lower latitudes are associated with higher annual precipitation ($r^2 = –0.87$, P-value < 0.001 between latitude and annual precipitation), and humid climates can support more vigorous vegetative growth. Similarly, the negative relationship between flowering time and annual mean temperature suggests that sorghum plants in hot (dry) weather (correlation between annual temperature and annual precipitation is –0.18, P-value < 0.001) will tend to flower early to avoid prolonged drought and high temperature on the field as an escape mechanism (Tuinstra *et al.*, 1997).

Redundancy analysis indicated that phenotypic variation explained more of the SNP variation than either of space and climatic factors. The small proportion of SNP variation explained by space, climatic factors, and phenotypic information in this study could be due to the limited number of accessions evaluated (< 210 for SNP-Phenotype-Space-Climate model and < 450 for SNP-Space-Climate model). A distinct geographical pattern of distribution was found for the alleles of NAM QTL (qSbUBL3.47 and qSbLBL2.63) associated with inflorescence genes *ramosa2* and *Sobic.002G247800* (sorghum paralog of *OsSPL14*) underlying branch length and panicle length, respectively. The minor allele (qSbUBL3.47 - G) associated with short upper branches was found at high frequency in the guinea (80%) and caudatum (78%) accessions (figure 6A). The major allele (qSbLBL2.63 - A) associated with long lower branches was found at high frequency (69%) in the Guinea accessions and at low frequency (19%) in the Caudatum accession (Figure 6B). Guinea sorghum types are known to have very long lower branches and short upper branches. This suggests that these QTL can facilitate marker-assisted selection for long lower branches and short upper branches in Nigerian Guinea types.

Genes associated with adaptive traits underlie GWAS

Evidence of genomic footprints of clinal adaptation was found in the genome of Nigerian sorghum germplasm. A major selective sweep was observed on chromosome six while other signals of selection (outliers) were found on all chromosomes (figure 3.6A). The genome-wide average negative Tajima's *D* observed in this study for the Nigerian compared to the positive Tajima's *D* observed in the global and West African germplasm suggests the possible action of positive selection that could have purged deleterious alleles from the Nigerian germplasm. GWAS QTL for panicle length did not colocalize with NAM QTL for panicle length (Thesis chapter 2). This could be due to reduced power as a result of small sample size and population structure since panicle morphology is highly structured in sorghum (Brown *et al.*, 2011). However, for flowering time, GWAS QTL co-localized with known sorghum flowering time genes *Ma1* and *Ma6* (S6_40491020 about 174 kb from *Ma1* and S6_799609 about 99 kb from *Ma6*). In addition, significant associations were only found for annual precipitation and mean precipitation in the driest quarter. One of the SNPs associated with precipitation in the driest quarter (S1_2374316) was found in the intragenic region of a *No Apical Meristem* gene (Cheng *et al.*, 2012). Since there is a spatial correlation between climatic factors and population structure in the data set, the small number of SNP-environment association may be due to the GWAS model, which accounted for population structure (Günther & Coop, 2013; Yoder *et al.*, 2014).

Conclusion

In this study, a representative sample of the Nigerian sorghum germplasm was genotyped to characterize its genetic diversity, and population structure and to identify genomic signatures of selection present in its genome. This study provides an assessment of the Nigerian germplasm genetic diversity. The absence of genetic differentiation between the Nigerian and West African germplasm suggests possible gene flow between Nigerian and other West African countries probably due to the previous exchange of genetic materials. In addition, this can also foster germplasm exchange between national breeding programs. Results further showed that some phenotypic traits have been shaped by local adaptation across the Nigerian agroclimatic gradient. In addition, the genome-wide pattern of nucleotide variation showed signals of footprints of selection and association signals underlying adaptive traits. However, few association signals identified suggest the reduced power of the GWAS due to population structure associated with the adaptive traits. The sorghum nested association mapping (NAM) population has demonstrated higher power for mapping such adaptive traits (Bouchet *et al.*, 2017).

Figures

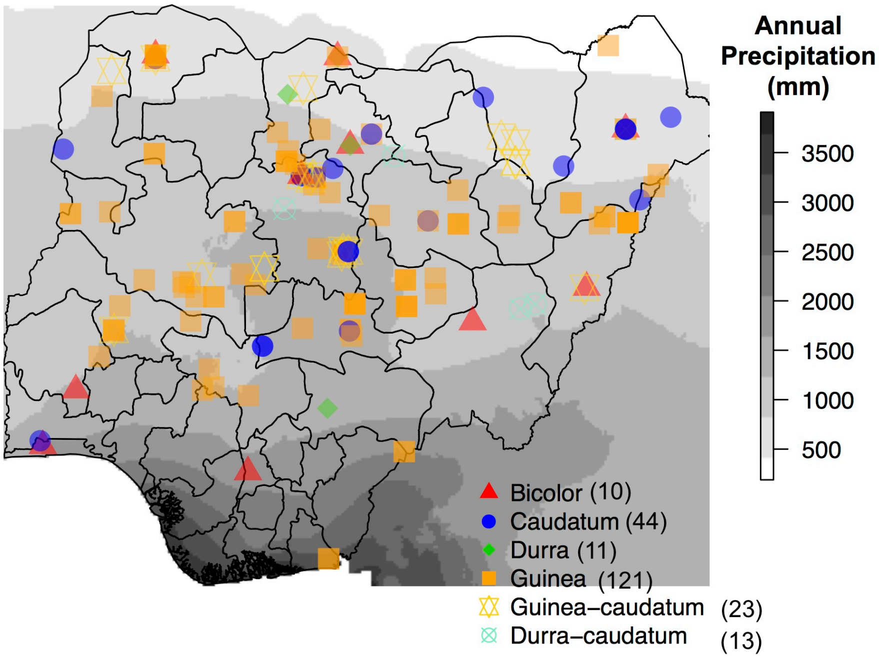

Figure 4.1: Distribution of sorghum accessions across precipitation zones in Nigeria.

Accessions are coded by botanical race, with number of accessions given in parentheses. Only accessions with known botanical race information are represented on this plot.

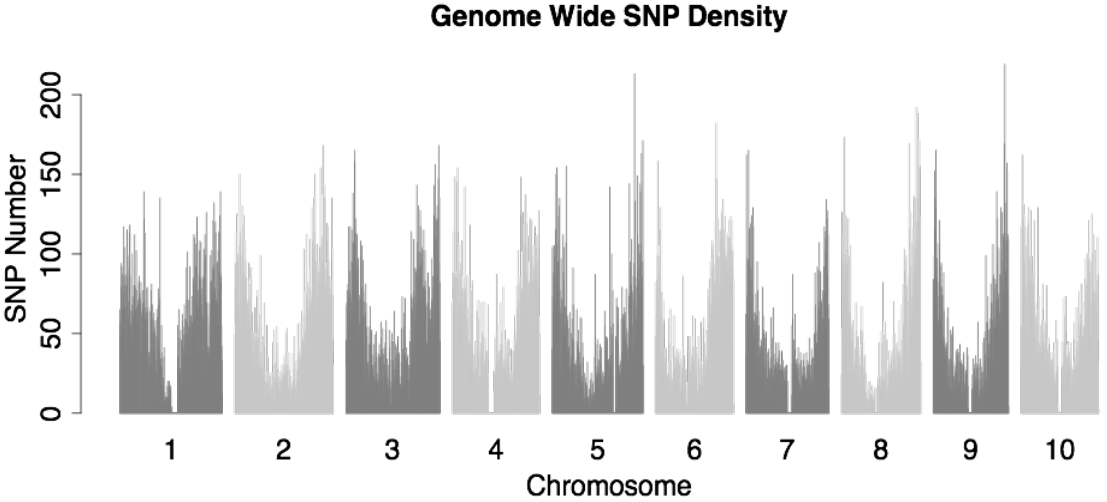

Figure 4.2: Genome wide single nucleotide polymorphisms (SNPs) density.

SNPs distribution across the genome (200 kb window) with high density in telomeric and sub-telomeric regions and reduced density in centromeric regions.

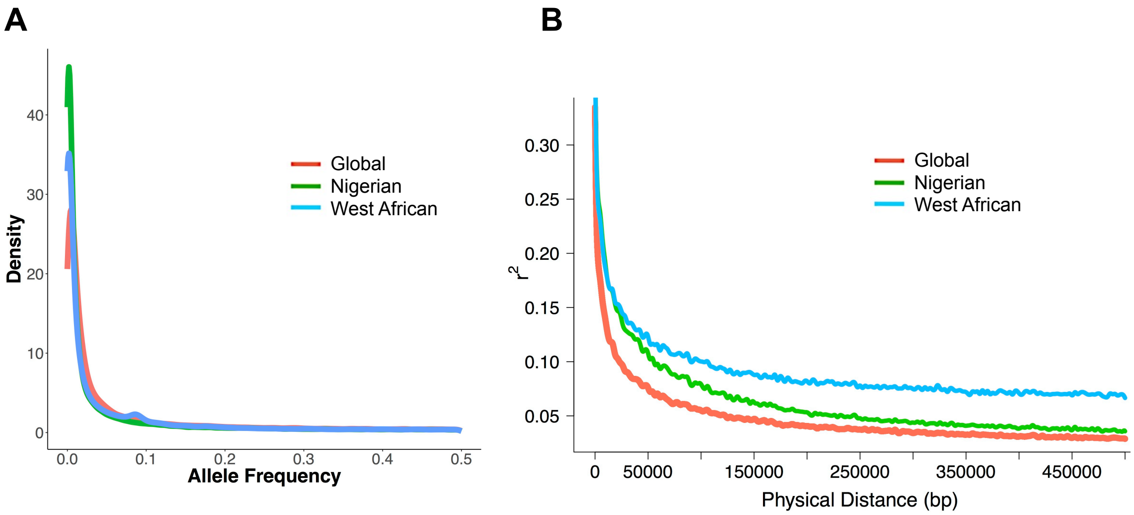

Figure 4.3: Minor allele frequencies and linkage disequilibrium decay.

(A) Minor allele frequency distribution of global reference (orange), Nigerian (green), and West Africa (WA) (medium blue). (B) Linkage disequilibrium curves for global reference (orange), Nigerian (green), and WA (medium blue).

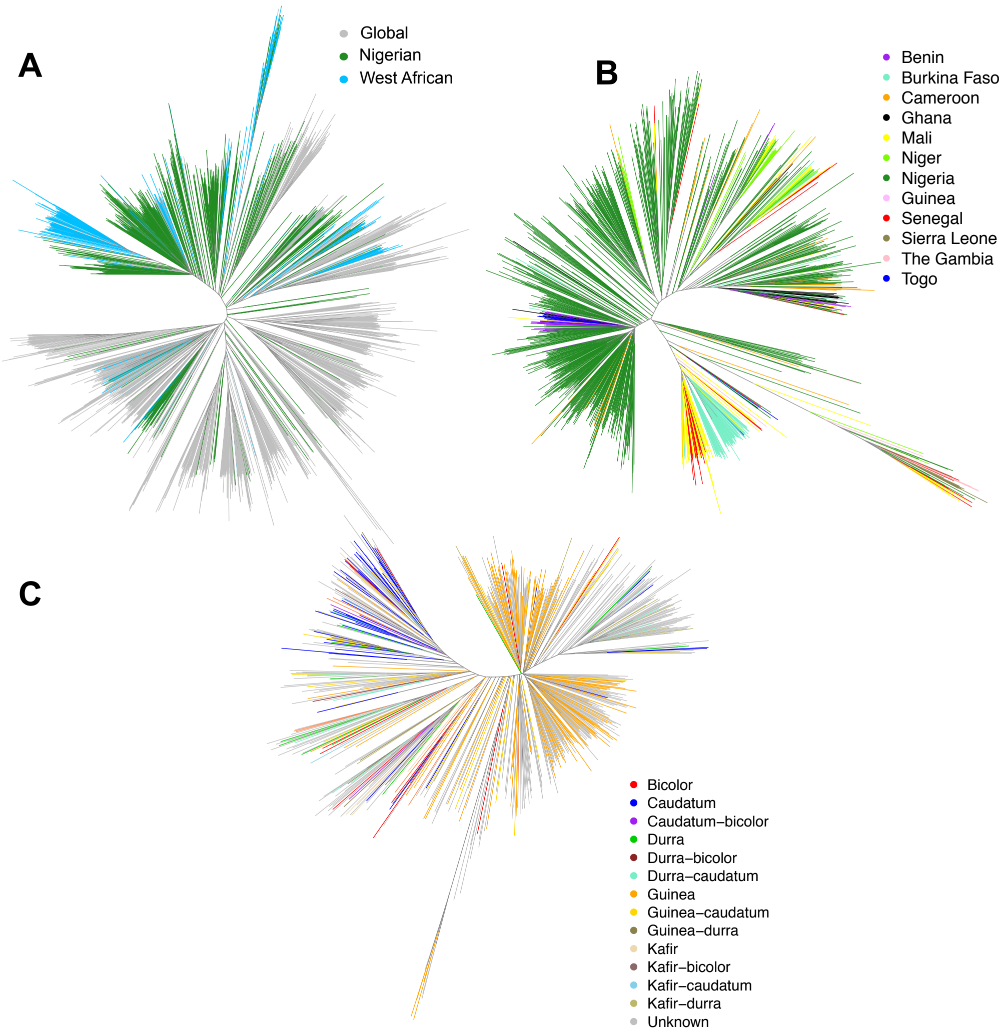

Figure 4.4: Genetic relatedness among in global, West African, and Nigerian accessions.

Neighbor joining analysis of (A) global germplasm (color-coded by panel), (B) West African germplasm (color-coded by country of origin), and (C) Nigerian germplasm (color-coded by botanical race).

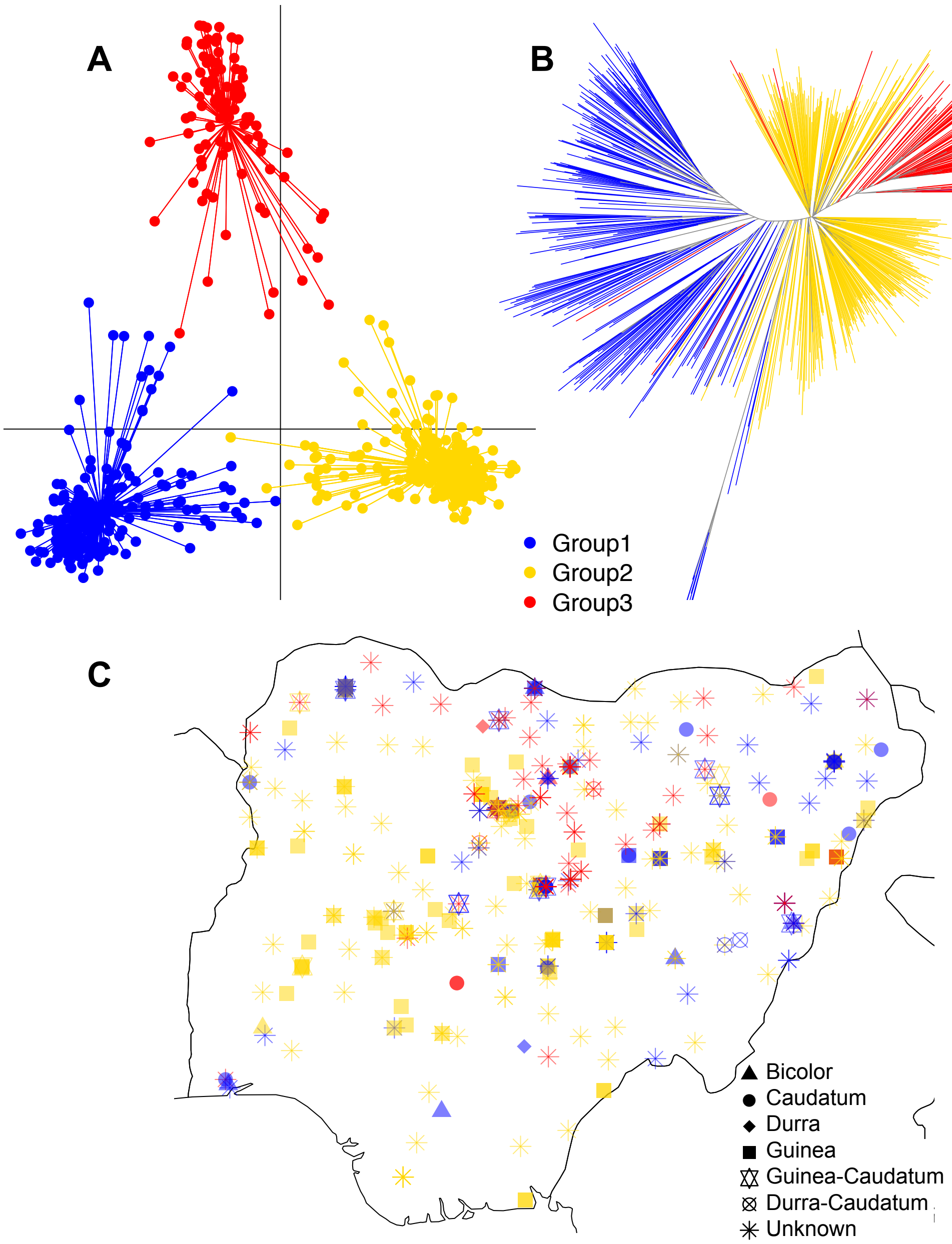

Figure 4.5: Discriminant analysis of principal components (DAPC) of the Nigerian germplasm.

(A) Discriminant analysis of principal component (DAPC) genetic groups, (B) neighbor joining tree color-coded based on DAPC groups, and distribution of georeferenced accessions across Nigerian geographical space. Groups are color-coded and shapes represent the sorghum races.

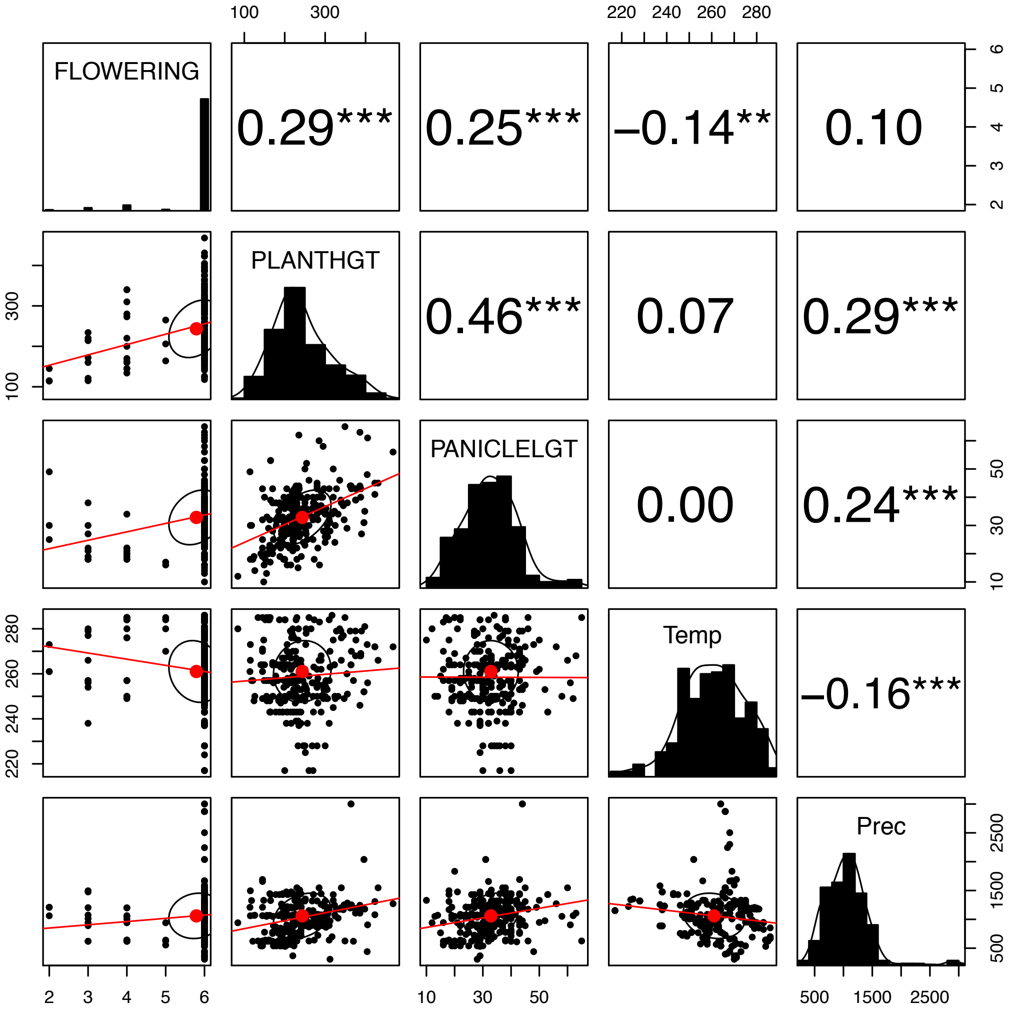

Figure 4.6: Pairwise correlation between traits and climatic variables.

Pearson correlation between flowering time (FLOWERING), plant height (PLANTHGT), panicle length (PANICLELGT), annual temperature (Temp), and annual precipitation (Prec) significant at 0.05, 0.01 and 0.001 (*, **, and ***)

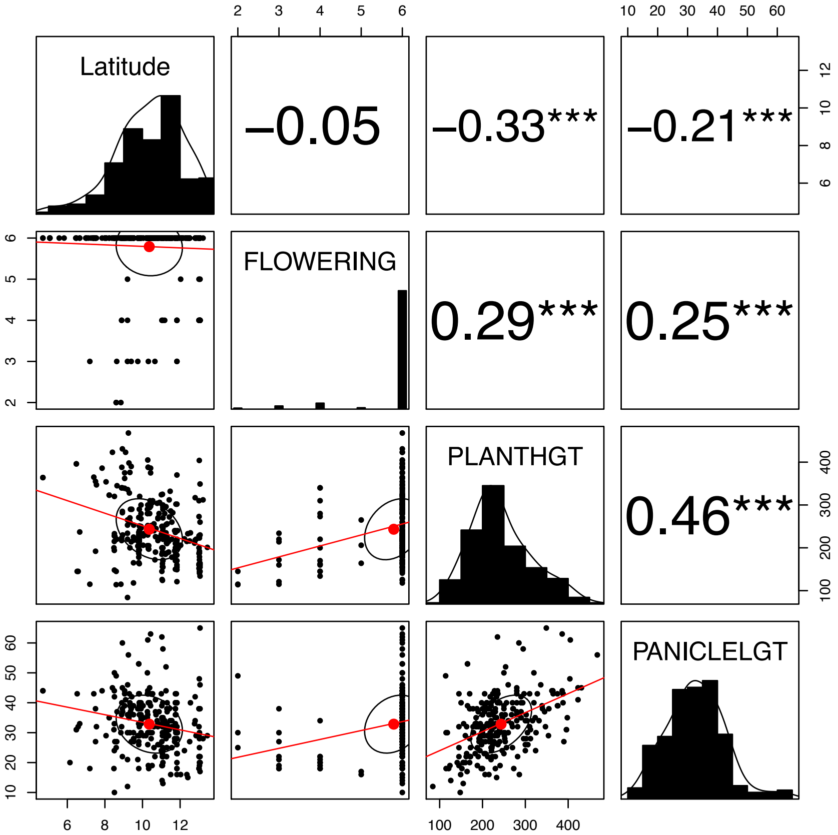

Figure 4.7: Pairwise correlation between geographical factor and traits.

Pearson correlation between latitude (Latitude), flowering time (FLOWERING), plant height (PLANTHGT), and panicle length (PANICLELGT) significant at 0.05, 0.01 and 0.001 (*, **, and ***).

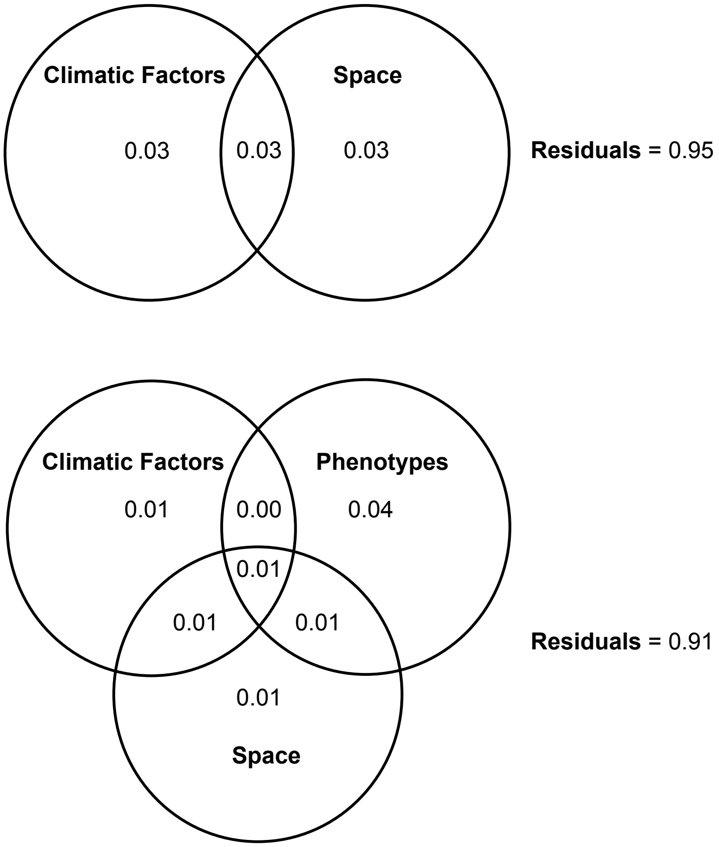

Figure 4.8: Proportion of SNP variation explained by climatic factors, space and phenotypes.

Multivariate redundancy analysis showing the proportion (0.0 to 1.0) of genotypic variation explained by climatic factors (temperature and precipitation), space (latitude and longitude), and phenotypes (flowering time, plant height, and panicle length).

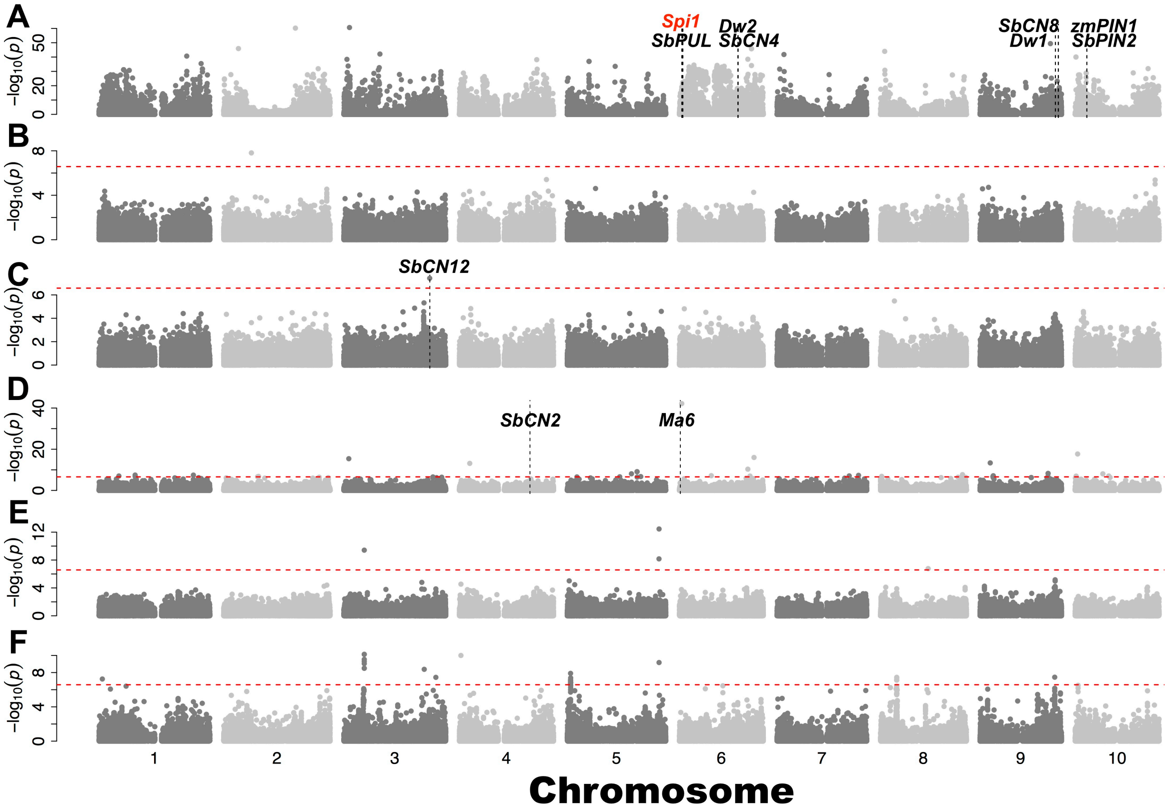

Figure 4.9: Genome-wide association studies of phenotype and climate in Nigerian germplasm.

Manhattan plot for (A) signatures of selection, showing outliers around *a priori* candidate genes. Broken lines shows the position of genes, text in red is the gene acronym for a paralog while texts in black are acronyms for gene orthologs in sorghum (B) panicle length, (C) plant height, (D) flowering time, (E) annual precipitation, and (F) precipitation in the driest quarter.

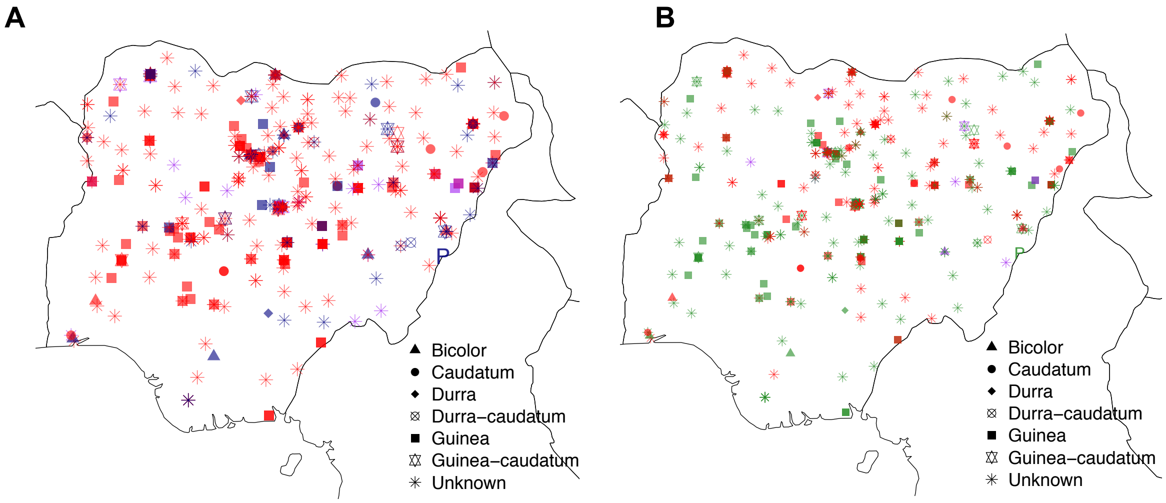

Figure 4.10: Geographic distribution of inflorescence QTL alleles.

(A) Geographic distribution of alleles of inflorescence upper branch length QTL (qSbUBL3.47 [S3_4750709]) associated with *Ramosa2* (Sobic.003G052900). (B) Geographic distribution of alleles of inflorescence lower branch length QTL (qSbLBL2.63 [S2_63576699]) associated with *OsSPL14* (Sobic.002G247800). Navy blue codes for 'C' allele, red codes for 'G' allele and Green codes for 'A' allele.

References

Barnaud A, Deu M, Garine E, Mckey D, Joly HI. **2007**. Local genetic diversity of sorghum in a village in northern Cameroon: structure and dynamics of landraces. *Theor Appl Genet* **114**: 237–248.

Bezançon G, Pham J-L, Deu M, Vigouroux Y, Sagnard F, Mariac C, Kapran I, Mamadou A, Gérard B, Ndjeunga J, *et al.* 2009. Changes in the diversity and geographic distribution of cultivated millet (Pennisetum glaucum (L.) R. Br.) and sorghum (Sorghum bicolor (L.) Moench) varieties in Niger between 1976 and 2003. *Genetic Resources and Crop Evolution* **56**: 223–236.

BGI-shenzhen. **2017**. BGI-shenzhen/PopLDdecay - Libraries.io.

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.

Brown PJ, Myles S, Kresovich S. **2011**. Genetic Support for Phenotype-based Racial Classification in Sorghum. *Crop Science* **51**: 224.

Browning BL, Browning SR. **2013**. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**: 459–71.

Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, Ersoz E, Flint-Garcia S, Garcia A, Glaubitz JC, *et al.* 2009. The Genetic Architecture of Maize. *Science* **325**: 714–718.

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**: 2449–63.

Cheng X, Peng J, Ma J, Tang Y, Chen R, Mysore KS, Wen J. **2012**. NO APICAL MERISTEM (MtNAM) regulates floral organ identity and lateral organ separation in Medicago truncatula. *New Phytologist* **195**: 71–84.

CNN. **2017**. Nigeria to surpass the US as world's 3rd most populous country - CNN.

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, Tang C, Stich B, Sikosek T, Effgen S, Josephs E, Schmitt J, Nordborg M, Koornneef M, de Meaux J. **2013**. Co-Variation between Seed Dormancy, Growth Rate and Flowering Time Changes with Latitude in Arabidopsis thaliana. *PLoS ONE* **8**.

Deu M, Sagnard F, Chantereau J, Calatayud C, Hérault D, Mariac C, Pham JL, 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.

Djè Y, Heuertz M, Lefèbvre C, Vekemans X. **2000**. Assessment of genetic diversity within and among germplasm accessions in cultivated sorghum using microsatellite markers. *TAG Theoretical and Applied Genetics* **100**: 918–925.

Ducrocq S, Madur D, Veyrieras J-B, Camus-Kulandaivelu L, Kloiber-Maitz M, Presterl T, Ouzunova M, Manicacci D, Charcosset A. **2008**. Key impact of Vgt1 on flowering time adaptation in maize: evidence from association mapping and ecogeographical information. *Genetics* **178**: 2433–7.

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. *PloS one* **6**: e19379.

Ezeaku IE, Gupta SC. **2004**. Development of sorghum populations for resistance to Striga hermonthica in the Nigerian Sudan Savanna. *African Journal of Biotechnology* **3**: 324–329.

Ezeaku IE, Gupta SC, Prabhakar VR. **1999**. Classification Of Sorghum Germplasm Accessions Using Multivariate Methods. *African Crop Science Journal* **7**: 97–108.

Feitosa Vasconcelos AC, Bonatti M, Schlindwein SL, D 'agostini R, Homem LR, Nelson R. **2013**. Landraces as an adaptation strategy to climate change for smallholders in Santa Catarina, Southern Brazil. *Land Use Policy* **34**: 250–254.

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 (NA Tinker, Ed.). *PLoS ONE* **9**: e90346.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, *et al.* **2012**. Phytozome: a comparative platform for green plant genomics. *Nucleic acids research* **40**: D1178-86.

Günther T, Coop G. **2013**. Robust identification of local adaptation from allele frequencies. *Genetics* **195**: 205–220.

Hamblin MT, Salas Fernandez MG, 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**: 1247–56.

Hancock AM, Brachi B, Faure N, Horton MW, Jarymowycz LB, Sperone FG, Toomajian C, Roux F, Bergelson J. **2011**. Adaptation to Climate Across the Arabidopsis thaliana Genome. *Science* **334**: 83–86.

Harlan J. **1992**. *Crops & Man* (G Peterson, S Baenziger, and R Dinauer, Eds.). Madison, WI, U.S.A: American Society of Agronomy.

van Heerwaarden J, Hufford MB, Ross-Ibarra J. **2012**. Historical genomics of North American maize. *Proceedings of the National Academy of Sciences* **109**: 12420–12425.

Hijmans RJ. **2016**. Geographic Data Analysis and Modeling. R package raster version

2.5-8.

Hufford MB, Xu X, van Heerwaarden J, Pyhäjärvi T, Chia J-M, Cartwright RA, Elshire RJ, Glaubitz JC, Guill KE, Kaeppler SM, *et al.* 2012. Comparative population genomics of maize domestication and improvement. *Nature Genetics* **44**: 808–811.

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

Kronholm I, Picó FX, Alonso-Blanco C, Goudet J, Meaux J de. **2012**. Genetic basis of adaptation in Arabidopsis thaliana: Local adaptation at the seed dormancy QTL DOG1. *Evolution* **66**: 2287–2302.

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–e1400218.

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

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.

Luu K, Bazin E, Blum MGB. **2017**. *pcadapt* : an R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources* **17**: 67–77.

Manzelli M, Pileri L, Lacerenza N, Benedettelli S, Vecchio V. **2007**. Genetic diversity assessment in Somali sorghum (Sorghum bicolor (L.) Moench) accessions using microsatellite markers. *Biodiversity and Conservation* **16**: 1715–1730.

Mercer KL, Perales HR, Wainwright JD. **2012**. Climate change and the transgenic adaptation strategy: Smallholder livelihoods, climate justice, and maize landraces in Mexico. *Global Environmental Change* **22**: 495–504.

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**: 453–8.

de Oliveira AC, Richter T, Bennetzen JL. **1996**. Regional and racial specificities in sorghum germplasm assessed with DNA markers. *Genome* **39**: 579–87.

Olsen KM, Caicedo AL, Polato N, Mcclung A, Mccouch S, Purugganan MD. **2006**. Selection Under Domestication: Evidence for a Sweep in the Rice Waxy Genomic Region. *Genetics* **173**: 975–983.

OYENUGA VA. **1967**. *Agriculture in Nigeria: An Introduction.* Rome: Food and Agriculture Organization of the United Nations.

Paradis E, Claude J, Strimmer K. **2004**. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics (Oxford, England)* **20**: 289–90.

Samis KE, Murren CJ, Bossdorf O, Donohue K, Fenster CB, Malmberg RL, Purugganan MD, Stinchcombe JR. **2012**. Longitudinal trends in climate drive flowering time clines in north american arabidopsis thaliana. *Ecology and Evolution* **2**: 1162–1180.

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.

Siol M, Wright SI, Barrett SCH. **2010**. The population genomics of plant adaptation. *New Phytologist* **188**: 313–332.

Soler C, Saidou A-A, Vi Cao Hamadou T, Pautasso M, Wencelius J, Joly HHI. **2013**. Correspondence between genetic structure and farmers' taxonomy – a case study from dry-season sorghum landraces in northern Cameroon. *Plant Genetic Resources* **11**: 36–49.

Sowunmi FA, Akintola JO. **2010**. Effect of Climatic Variability on Maize Production in Nigeria. *Research Journal of Environmental and Earth Sciences* **2**: 19–30.

Tesso T, Kapran I, Grenier C, Snow A, Sweeney P, Pedersen J, Marx D, Bothma G, Ejeta G. **2008**. The Potential for Crop-to-Wild Gene Flow in Sorghum in Ethiopia and Niger: A Geographic Survey. *Crop Science* **48**: 1425.

Thurber CS, Ma JM, Higgins RH, Brown PJ. **2013**. Retrospective genomic analysis of sorghum adaptation to temperate-zone grain production. *Genome Biology* **14**: R68.

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.

Umina PA, Weeks AR, Kearney MR, McKechnie SW, Hoffmann AA, Sperone FG, Toomajian C, Roux F, Bergelson J. **2005**. A rapid shift in a classic clinal pattern in Drosophila reflecting climate change. *Science (New York, N.Y.)* **308**: 691–3.

Vigouroux Y, Mariac C, de Mita S, Pham JL, G??rard B, Kapran I, Sagnard F, Deu M, Chantereau J, Ali A, *et al.* **2011**. Selection for earlier flowering crop associated with climatic variations in the Sahel. *PLoS ONE* **6**: 1–9.

Worldclim.org. WorldClim 1.4: Current conditions (~1960-1990) | WorldClim - Global Climate Data.

WSP. **2017**. Sorghum | World Sorghum Production 2017/2018 https://www.worldsorghumproduction.com/previous-month.asp.

Yoder JB, Stanton-Geddes J, Zhou P, Briskine R, Young ND, Tiffin P. **2014**. Genomic Signature of Adaptation to Climate in Medicago truncatula. *Genetics* **196**: 1263–1275.

Zeven AC. **1998**. Landraces: A review of definitions and classifications. *Euphytica* **104**: 127–139.

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.

Zhen Y, Ungerer MC. **2008**. Clinal variation in freezing tolerance among natural accessions of Arabidopsis thaliana. *New Phytologist* **177**: 419–427.

Appendix A - NAM Inflorescence QTL

Marker	QTL	Trait	Gene Name	Relationship	Similarity	Sorghum Gene ID
S7_59751994	SbLBL7.59	LBL_GWAS	sparse inflorescence1 (spi1)	Paralog	55.1	Sobic.007G163200
S7_59751994	qSbLBL7.59	LBL_GWAS	Dwarf3(Dw3)	Ortholog	NA	Sobic.007G163800
S2_63576699	qSbLBL2.63	LBL_JL	OsSPL14	Paralog	47.2	Sobic.002G247800
S3_4450819	qSbLBL3.44	LBL_JL	Barren Inflorescence 2 (bif2)	Paralog	43.9	Sobic.003G048700
S1_21494247	qSbLBL1.21	LBL_JL	Fasciated ear 2 (Fea2)	Paralog	37	Sobic.001G224000
S2_63576699	qSbLBL2.63	LBL_JL	Fasciated ear4 (Fea4)	Paralog	48.4	Sobic.002G247300
S3_8666433	qSbLBL3.86	LBL_JL	BARRENSTALK1 (BA1)	Paralog	21.5	Sobic.003G099000
S1_6775611	qSbLBL1.67	LBL_JL	LEAFY HULL STERILE1 (LHS1)/OsMADS1	Paralog	59.5	Sobic.001G086400
S1_6775611	qSbLBL1.67	LBL_JL	Panicle Phytomer2 (PAP2)	Ortholog	90.2	Sobic.001G086400
S1_6775611	qSbLBL1.67	LBL_JL	Panicle Phytomer2 (PAP2)	Paralog	69.1	Sobic.001G086400
S1_6775611	qSbLBL1.67	LBL_JL	Sepetalla	Paralog	68.4	Sobic.001G086400
S10_51874918	qSbLBL10.51	LBL_JL	BARREN INFLORESCENCE 1	Paralog	22.9	Sobic.010G180600
S2_63475769	qSbLBL2.63	LBL_NJL	Fasciated ear4 (Fea4)	Paralog	48.4	Sobic.002G247300
S7_57067166	qSbLBL7.57	LBL_NJL	Fasciated ear 2 (Fea2)	Paralog	34.4	Sobic.007G141700
S10_51414143	qSbLBL10.51	LBL_NJL	Fasciated ear 2 (Fea2)	Paralog	36.7	Sobic.010G177300
S10_51414143	qSbLBL10.51	LBL_NJL	THICK TASSEL DWARF1/CLAVATA1	Paralog	49.2	Sobic.010G177300
S2_63475769	qSbLBL2.63	LBL_NJL	OsSPL14	Paralog	47.2	Sobic.010G177300
S7_59953003	qSbLBL7.59	LBL_NJL	Dwarf3(Dw3)	Ortholog	NA	
S1_63043663	qSbLBL1.63	LBL_NJL	DENSE AND ERECT PANICLE (DEP1)	Paralog	10.1	Sobic.001G341700
S2_60856616	qSbRD2.60	RD_JL	DENSE AND ERECT PANICLE (DEP1)	Paralog	15	Sobic.002G216600
S3_73412473	qSbRD3.73	RD_JL	Fasciated ear 2 (Fea2)	Paralog	35.2	Sobic.003G432000
S1_4131827	qSbRD1.41	RD_JL	BARREN INFLORESCENCE 1	Paralog	21.4	Sobic.001G056100
S1_4131827	qSbRD1.41	RD_JL	BARREN INFLORESCENCE 4	Paralog	27.4	Sobic.001G056100
S9_57369245	qSbRD9.57	RD_JL	WUSCHEL-related homeobox 1A	Paralog	13.8	Sobic.009G233000
S1_58623035	qSbRD1.58	RD_JL	Ramosa3 (ra3)	Paralog	55.5	Sobic.001G303900
S6_57581969	qSbRD6.57	RD_NJL	Fasciated ear4 (Fea4)	Paralog	42.4	Sobic.006G233500
S7_57230541	qSbRD7.57	RD_NJL	Fasciated ear 2 (Fea2)	Paralog	34.4	Sobic.007G141700
S3_61838244	qSbRD3.61	RD_NJL	sparse inflorescence1 (spi1)	Paralog	70.8	Sobic.003G286500
S7_59751994	qSbRL7.59	RL_GWAS	sparse inflorescence1 (spi1)	Paralog	55.1	Sobic.007G163200
S7_59751994	qSbRL7.59	RL_GWAS	Dwarf3(Dw3)	Ortholog	NA	Sobic.007G163800

S7_59751994	qSbRL7.59	RL_JL	sparse inflorescence1 (spi1)	Paralog	55.1	Sobic.007G163200
S1_78453360	qSbRL1.78	RL_JL	BARRENSTALK1 (BA1)	Paralog	23.7	Sobic.001G518900
S7_59751994	qSbRL7.59	RL_JL	Dwarf3(Dw3)	Ortholog	NA	Sobic.007G163800
S8_58771392	qSbRL8.58	RL_JL	BARRENSTALK1 (BA1)	Paralog	NA	
S8_58771392	qSbRL8.58	RL_JL	BARRENSTALK1 (BA1)	Paralog	NA	Sobic.008G154100
S4_58277149	qSbRL4.58	RL_JL	Ramosa3 (ra3)	Paralog	58.5	Sobic.004G232900
S1_63011488	qSbRL1.63	RL_JL	DENSE AND ERECT PANICLE (DEP1)	Paralog	10.1	Sobic.001G341700
S1_75419531	qSbRL1.75	RL_JL	Branched silkless1 (bd1)	Paralog	14	Sobic.001G481400
S1_21565786	qSbRL1.21	RL_JL	Fasciated ear 2 (Fea2)	Paralog	37	Sobic.001G224000
S8_58771392	qSbRL8.58	RL_JL	BARRENSTALK1 (BA1)	Paralog	17.4	Sobic.008G154000
S8_58771392	qSbRL8.58	RL_JL	BARRENSTALK1 (BA1)	Paralog	NA	Sobic.008G153900
S8_58771392	qSbRL8.58	RL_JL	BARREN INFLORESCENCE 1	Paralog	26.2	Sobic.008G153900
S8_58771392	qSbRL8.58	RL_JL	BARREN INFLORESCENCE 4	Paralog	27.8	Sobic.008G153900
S7_56004399	qSbRL7.56	RL_NJL	Tunicate1	Paralog	42	Sobic.007G135301
S3_69363350	qSbRL3.69	RL_NJL	LEAFY HULL STERILE1 (LHS1)/OsMADS1	Paralog	49.8	Sobic.003G381100
S3_69363350	qSbRL3.69	RL_NJL	OsMADS58	Paralog	54.2	Sobic.003G381100
S3_69363350	qSbRL3.69	RL_NJL	Sepetalla	Paralog	52	Sobic.003G381100
S7_59953003	qSbRL7.59	RL_NJL	Dwarf3(Dw3)	Ortholog	NA	Sobic.007G163800
S3_4750709	qSbUBL3.47	UBL_GWAS	Ramosa2 (ra2)	Ortholog	82.8	Sobic.003G052900
S3_4750709	qSbUBL3.47	UBL_GWAS	Fasciated ear 2 (Fea2)	Paralog	36.2	Sobic.003G052100
S3_69496539	qSbUBL3.69	UBL_JL	LEAFY HULL STERILE1 (LHS1)/OsMADS1	Paralog	49.8	Sobic.003G381100
S3_69496539	qSbUBL3.69	UBL_JL	OsMADS58	Paralog	54.2	
S3_69496539	qSbUBL3.69	UBL_JL	Sepetalla	Paralog	52	Sobic.003G381100
S3_4750709	qSbUBL3.47	UBL_JL	Ramosa2 (ra2)	Ortholog	82.8	Sobic.003G052900
S3_4750709	qSbUBL3.47	UBL_JL	Fasciated ear 2 (Fea2)	Paralog	36.2	Sobic.003G052100
S3_73583203	qSbUBL3.73	UBL_JL	Fasciated ear 2 (Fea2)	Paralog	35.2	Sobic.003G432000
S3_4757321	qSbUBL3.47	UBL_NJL	Ramosa2 (ra2)	Ortholog	82.8	Sobic.003G052900
S3_4757321	qSbUBL3.47	UBL_NJL	Fasciated ear 2 (Fea2)	Paralog	36.2	Sobic.003G052100

Appendix B - NAM Vegetative QTL

Marker	QTL	Trait	Gene.Name	Relationship	Similarity	Sorghum Gene ID
S6_51325786	qSbHGT6.51	HGT	BAD1	Paralog	11.6	Sobic.006G154000
S7_59611315	qSbHGT7.59	HGT	Sparse inflorescence1 (Spi1)	Paralog	55.1	Sobic.007G163200
S9_57215490	qSbHGT9.57	HGT	narrow sheath1	Paralog	14.1	Sobic.009G233000
S7_59787744	qSbHGT7.59	HGT	Dwarf3 (Dw3)	Ortholog	NA	Sobic.007G163800
S7_59787744	qSbHGT7.59	HGT	sparse inflorescence1 (spi1)	Paralog	55.1	Sobic.007G163200
S6_42798327	qSbHGT6.42	HGT	Dwarf2 (Dw2)	Ortholog	NA	Sobic.006G067700
S6_42798327	qSbHGT6.42	HGT	SbCN4	Ortholog	NA	Sobic.006G068300
S9_57065264	qSbHGT9.57	HGT	Dwarf1 (Dw1)	Ortholog	NA	Sobic.009G229800
S3_71464034	qSbFLT3.71	FLT	BAD1	Paralog	17.4	Sobic.003G408400
S6_799602	qSbFLT6.79	FLT	Ma6	Ortholog	NA	Sobic.006G004400
S3_62717707	qSbFLT3.62	FLT	SbCN12	Ortholog	NA	Sobic.003G295300
S6_799654	qSbFLT6.79	FLT	Ma6	Ortholog	NA	Sobic.006G004400
S7_5137535	qSbSTM7.51	STM	ROUGH SHEATH2	Paralog	11.4	Sobic.007G050400
S3_63584041	qSbSTM3.63	STM	BAD1	Paralog	17.4	Sobic.003G305000
S3_63584041	qSbSTM3.63	STM	BLADE-ON-PETIOLE1/2	Paralog	39	Sobic.003G308700
S6_51419656	qSbLET6.51	LET	BAD1	Paralog	11.6	Sobic.006G154000
S2_3403756	qSbLET2.34	LET	BAD1	Paralog	18.5	Sobic.002G035500