Genomic tools for exploiting germplasm resources to improve grain attributes in sorghum: A case of Ethiopian sorghum germplasm collection

by

Yemane Girma Belaineh

B.S., Haramaya University (formerly Alemaya University), Ethiopia, 2004
M.S., University of Agricultural Sciences, Dharwad, India, 2009

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

2023

Abstract

Sorghum is a primary source of diet for millions of people living in the semi-arid regions of Sub-Saharan Africa and Asia. Due to its immense resilience, sorghum stands as the crop of choice in the face of climate change that has already been causing widespread crop failures. However, the low nutritional quality of sorghum has negatively impacted its use and marketability relative to other cereals. Given the vast untapped germplasm resources for the species, opportunities exist to exploit beneficial alleles that may be of value to tackle challenges related to sorghum production and utilization. The current work is focused on exploring germplasm resources from one of the most significant sources of diversity, Ethiopia, to lay the scientific basis for genetic improvement of sorghum nutritional traits with emphasis on protein and the role of grain physicochemical attributes on adaptation behavior of the species. The work is presented in four chapters. The first chapter deals with a review of background information on the nutritional attributes of cereals emphasizing on challenges and opportunities for improving protein content; the second part investigates the pattern of adaptation of sorghum across Ethiopia's diverse agroecology in view of bioclimatic factors vis-a-vis grain physicochemical attributes and genomic profile; the third chapter explores the power of genomics for mining germplasm resources in gene banks; the last chapter focuses on the impact of grain pre-treatment on bio-availability of proteins from a fermented sorghum food product.

In the second chapter, after the background review, the hypothesis that environmental factors shape sorghum grain attributes was tested using more than 1500 Ethiopian landraces. We utilized phenotype-environment and genome-environment associations to support the thesis. The phenotype-environment association supports the hypothesis that tannin presence, grain weight, kernel hardness, and panicle compactness are all associated with historic precipitation gradient. The correlation pattern revealed by principal component analysis fits the expectation that grain attributes that favor grain-related diseases, such as compact panicles, were mainly concentrated in drier areas. In contrast, traits like tannin presence and loose panicle dominate high precipitation areas. Moreover, landraces from low rainfall regions were susceptible to grain mold suggesting the need to incorporate resistance when materials from dry regions are used as breeding parent for developing varieties for high precipitation areas. Genome-environment association also revealed

the importance of polyphenols for the adaptation of sorghum. Moreover, the genomic loci attributed to historical population structure were correlated with precipitation and temperature gradients. The study suggests that sorghum improvement endeavors targeting grain attributes should also consider the climatic condition of the target environments. Likewise, germplasm originating from high precipitation areas may be utilized as donors of resistance genes to various grain diseases

The third section investigates the potential of genomic selection (GS) in germplasm improvement. The study utilized grain-related and phenological data from Ethiopian sorghum core collection. Low to moderate prediction and validation accuracies were observed for the traits and increasing training size increased prediction accuracy. The focused identification of germplasm sampling (FIGS) approach, which had been proved successful in increasing the success rate in identifying rare alleles from large germplasm collections, was also evaluated for its complementarity with GS. Grain weight was utilized as a proxy for assessing the approach. Sampling using the FIGS-based approach changed population parameters relative to the base population. Genomic prediction on a reference population sampled using FIGS based approach had smaller validation accuracy and selection differential than randomly reconstituted reference populations. Modifying the FIGS sampling strategy by incorporating a few individuals from the opposite end of the FIGS predicted environment improved the overall performance of the system.

The last chapter investigated the importance of pre-processing method to improve protein digestibility, a critical constraint in sorghum. This was conducted using four preprocessing methods on four selected varieties of sorghum varying in grain quality attributes. The result showed significant pre-processing and variety interaction effects in protein digestibility of fermented and cooked sorghum food samples, implying that varietal selection should target a specific pre-processing method. Sprouting, one of the pre-treatment methods studied, improved overall grain protein digestibility. Genotypes with inherently improved protein content and in-vitro protein digestibility when subjected to appropriate milling and pre-processing treatment can significantly enhance protein availability from fermented sorghum foods.

In conclusion, understanding the adaptation history and the target end-user application is crucial for improving sorghum grain quality and nutritional traits. The information generated on

the grain attributes and the genomic selection pipeline for the FIGS approach has promising potential to accelerate the development of nutritionally improved and locally adapted varieties.

Genomic tools for exploiting germplasm resources to improve grain attributes in sorghum: A case of Ethiopian sorghum germplasm collection

by

Yemane Girma Belaineh

B.S., Haramaya University (formerly Alemaya University), Ethiopia, 2004
M.S., University of Agricultural Sciences, Dharwad, India, 2009

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

2023

Approved by:
Major Professor
Tesfaye Tesso

Copyright

Abstract

Sorghum is a primary source of diet for millions of people living in the semi-arid regions of Sub-Saharan Africa and Asia. Due to its immense resilience, sorghum stands as the crop of choice in the face of climate change that has already been causing widespread crop failures. However, the low nutritional quality of sorghum has negatively impacted its use and marketability relative to other cereals. Given the vast untapped germplasm resources for the species, opportunities exist to exploit beneficial alleles that may be of value to tackle challenges related to sorghum production and utilization. The current work is focused on exploring germplasm resources from one of the most significant sources of diversity, Ethiopia, to lay the scientific basis for genetic improvement of sorghum nutritional traits with emphasis on protein and the role of grain physicochemical attributes on adaptation behavior of the species. The work is presented in four chapters. The first chapter deals with a review of background information on the nutritional attributes of cereals emphasizing on challenges and opportunities for improving protein content; the second part investigates the pattern of adaptation of sorghum across Ethiopia's diverse agroecology in view of bioclimatic factors vis-a-vis grain physicochemical attributes and genomic profile; the third chapter explores the power of genomics for mining germplasm resources in gene banks; the last chapter focuses on the impact of grain pre-treatment on bio-availability of proteins from a fermented sorghum food product.

In the second chapter, after the background review, the hypothesis that environmental factors shape sorghum grain attributes was tested using more than 1500 Ethiopian landraces. We utilized phenotype-environment and genome-environment associations to support the thesis. The phenotype-environment association supports the hypothesis that tannin presence, grain weight, kernel hardness, and panicle compactness are all associated with historic precipitation gradient. The correlation pattern revealed by principal component analysis fits the expectation that grain attributes that favor grain-related diseases, such as compact panicles, were mainly concentrated in drier areas. In contrast, traits like tannin presence and loose panicle dominate high precipitation areas. Moreover, landraces from low rainfall regions were susceptible to grain mold suggesting the need to incorporate resistance when materials from dry regions are used as breeding parent for developing varieties for high precipitation areas. Genome-environment association also revealed

the importance of polyphenols for the adaptation of sorghum. Moreover, the genomic loci attributed to historical population structure were correlated with precipitation and temperature gradients. The study suggests that sorghum improvement endeavors targeting grain attributes should also consider the climatic condition of the target environments. Likewise, germplasm originating from high precipitation areas may be utilized as donors of resistance genes to various grain diseases

The third section investigates the potential of genomic selection (GS) in germplasm improvement. The study utilized grain-related and phenological data from Ethiopian sorghum core collection. Low to moderate prediction and validation accuracies were observed for the traits and increasing training size increased prediction accuracy. The focused identification of germplasm sampling (FIGS) approach, which had been proved successful in increasing the success rate in identifying rare alleles from large germplasm collections, was also evaluated for its complementarity with GS. Grain weight was utilized as a proxy for assessing the approach. Sampling using the FIGS-based approach changed population parameters relative to the base population. Genomic prediction on a reference population sampled using FIGS based approach had smaller validation accuracy and selection differential than randomly reconstituted reference populations. Modifying the FIGS sampling strategy by incorporating a few individuals from the opposite end of the FIGS predicted environment improved the overall performance of the system.

The last chapter investigated the importance of pre-processing method to improve protein digestibility, a critical constraint in sorghum. This was conducted using four preprocessing methods on four selected varieties of sorghum varying in grain quality attributes. The result showed significant pre-processing and variety interaction effects in protein digestibility of fermented and cooked sorghum food samples, implying that varietal selection should target a specific pre-processing method. Sprouting, one of the pre-treatment methods studied, improved overall grain protein digestibility. Genotypes with inherently improved protein content and in-vitro protein digestibility when subjected to appropriate milling and pre-processing treatment can significantly enhance protein availability from fermented sorghum foods.

In conclusion, understanding the adaptation history and the target end-user application is crucial for improving sorghum grain quality and nutritional traits. The information generated on

the grain attributes and the genomic selection pipeline for the FIGS approach has promising potential to accelerate the development of nutritionally improved and locally adapted varieties.

.

Table of Contents

List of Figures

List of Tables

Acknowledgments

"Coming together is a beginning, staying together is progress, and working together is a success." –Henry Ford

This work would not have been successful without the support of many. I am deeply thankful to my adviser Dr. Tesfaye Tesso for his availability, support, encouragement, and constructive critique. I would also like to thank Dr. Scott Bean, who availed working space at USDA, Manhattan, and helped me with food processing, for which I was a novice. I am also thankful to my committee members, Dr. Ramasamy Perumal and Dr. Sanzhen Liu, for their crucial remarks and help; thank you! I would also like to thank Dr. Geoffrey Morris for teaching me genomic tools and research methodologies essential for my study. I am also indebted to Dr. Fonscea and Dr Habtamu Ayalew for their kind assistance on critiquing part of the manuscript.

I want to forward my gratitude to Dr. Shantha Peiris, who helped me calibrate NIRS spectral data. I am thankful to Dr. Getachew Ayana, Mr. Amare Siyoum, Dr. Habte Nida, Dr. Taye Tadesse, Mr. Daniel Nadew, Dr. Firew Mekbib, Ms. Alemenesh Bekele, Mr Negese Tujuba, Dr Gezahegn Girma and Ms Dultu Talili for their support at different stages of my work.

I am thankful to Sorghum and Millet Innovation Lab for the financial support. I am also thankful to the Computing and Information Sciences Department, Kansas State University, for availing the Beocat computing cluster for storing and analyzing genomic data.

Thanks to my lab mates, Dr. Dereje Dugassa, Mr. Diriba Chere, Mr. Kinde Nouh, and Mr. Daniel Hopper, for the support and good moments. Thank you! I express my deepest gratitude to Ms. Kira Everhart, Ms. Nancy Williams, and Ms. Karlene Teske for your help during my stay.

Dedication

I dedicate this work to my beloved wife, Kidist Dawit, for her unwavering support and sacrifice.

Chapter 1 - Breeding for improved grain protein concentration in cereals: challenges and opportunities

Introduction

Cereals, including wheat, maize, rice, barley, and sorghum, are among the major food/feed grains throughout the world. They are particularly rich in starch and serve as the primary source of calories in both animal feed and human food. In 2020, 736 million ha of land was estimated to be dedicated to cereal crop production worldwide, and 2.99 billion tons of grain were produced. Pulses, the major protein crops, covered 93 million hectares and produced 89 million tons of grains (FAOSTAT, 2020). The dominance of cereals over other crops appears to be due to their wide agroclimatic adaptability that ranges from the hot, dry tropical environment to cold zones of the world and their efficiency in dry biomass production. Wheat and rice are mainly consumed as food, where wheat is mainly used for baked products such as bread and pasta, while rice is consumed directly wet cooked. In the developing world, coarse cereals such as sorghum and maize are primarily utilized as human food, while they are mainly grown as animal feed in the developed world (Zhou, 2009). Besides their direct use food and feed grain, cereals are also essential raw materials in processed foods and chemical industries. Maize starch is used as raw material for making various industrial products, including corn syrup, bioethanol production, biodegradable packaging material, and adhesives. The wet and dry milling by-products such as oil and protein concentrate are also essential economic products. Barley malt is another critical industrial product used in the brewery industry. The by-product of malting, brewer's spent grain, is rich in protein and other nutrients and is used as a good source of protein in animal feeding (Papageorgiou and Skendi, 2018).

Cereals as dietary protein sources

Protein is the second major component of cereal grains. Cereal protein is the major dietary source of protein among smallholder communities in developing countries where access to animal protein is limited. There are marked differences in protein content both between and within cereal

species (Table 1-1). According to FAO (2020) estimates, the world protein supply from cereals (39%) was almost equal to the supply from animal protein sources (also 39%), while the share of pulses was only around 5% FAOSTAT (2020). The importance of cereals as a protein source becomes even more evident in the developing world, where cereals account for 51% of the total protein supply in the diets of the people (FAOSTAT, 2020) (Figure 1-1). Cereal protein utilization is not uniform across societies either, with populations at the bottom of the socio-economic ladder increasingly relying on cereal proteins (Haileselassie et al., 2020).

The demand for protein is projected to increase due to population growth and socio-economic changes favoring extra protein consumption. Moreover, the boom in income status of the developing world's middle-class community is expected to boost animal protein consumption, despite the fact that animal protein production is less efficient and requires more land and energy (Henchion et al., 2017). Thus, crops will continue to be the ultimate source of protein for direct or indirect consumption through animal products or other processed products. As an example, the global market for wheat protein in the form of gluten, protein isolate, and protein hydrolysate is estimated at $1.11 billion in 2018 and is projected to increase by 12% by the year 2026 (Verified Market Research, 2019). The alcohol-soluble prolamins of wheat, sorghum, maize, and barley have potential industrial and pharmaceutical use as biodegradable plastics, biofilms, and drug delivery capsules (Taylor et al., 2013) that their global demand is on the rise.

The grain protein content is often used as a proxy to determine the food product quality and pricing of grains (Dexter et al., 1994). In wheat, gluten proteins, the major storage proteins contributing more than 80% of the grain protein, are important determinants of the overall bread-making quality. The grain protein content is directly correlated with the quantity of gluten protein as the increment of protein content usually disproportionately increases gluten fraction than other non-gluten fractions (Arendt et al., 2008). However, the quantity of high molecular polymeric glutens- glutenin polypeptides and the relative proportions of gliadins-gluten monomers determine different attributes of dough and the overall bread-making quality of wheat cultivars (Dhaka and Khatkar, 2015; Li et al., 2020). While molecular markers linked to the trait are proposed to better explain the quality differences among bread wheat (Cato and Mullan, 2020), elevators across North America use grain protein content as a practical proxy for determining premiums and discounts (Bekkerman, 2021). Similarly, protein content is positively correlated to pasta cooking quality of

durum wheat which include firmness, resistance to overcooking, and reduced stickiness (Sissons et al., 2021). In malting barley, protein has mixed importance as barley harvest incurs a penalty for protein content above and below a certain maximum and minimum thresholds (Dykha et al., 2021).

Grain crops like sorghum may also benefit from increased grain protein content to expand their niche market in the developed world. There is increasing acceptability of sorghum as a non-gluten (Fenster, 2003) food alternative rich in health-promoting phytochemicals (Awika and Rooney, 2004). Moreover, increased grain protein content means higher protein intake to consumers of the grain, be it in the feedlot in the developed world or as food for subsistence farmers in impoverished countries. As grain protein plays a vital role in health and food security, it is important to review and determine grain protein content, challenges faced in improving grain protein, and opportunities as solutions to the challenges.

Methods for evaluating grain protein content

The grain protein content is often expressed as the percentage of protein to grain constituents, expressed on a weight-by-weight basis. The building blocks of proteins, amino acids, are composed of nitrogen, oxygen, and hydrogen. Two of the twenty amino acids are also sulfur-containing. The primary techniques used for estimating protein aim to quantify the concentration of these building blocks in a sample.

N Based techniques

Crude protein in grains is mainly estimated indirectly from grain N content. Two methods, i.e., Kjeldahl and Dumas, are widely used for protein estimation. The Kjeldahl technique (and its modification) is widely used in grain analysis and involves two steps: the first step is strong acid-mediated digestion of the sample to convert sample N to ammonia salt, and the second step is distilling ammonia and quantifying the released ammonia (Miller and Houghton, 1945). The Dumas method involves three steps: first, the complete combustion of organic N to different N oxides, reducing the N oxides to molecular N (N2), and quantifying N using thermal conductivity. These two methods are standard protein determination in cereals (Simonne et al., 1997; Moore et al., 2010). The N quantified in a sample is converted to protein using a factor assumed to represent

the N percentage in proteins. The default conversion factor is 6.25. This constant implies that the N quantified is exclusively from amino-N of proteins, and the contribution of non-protein organic or inorganic N from the sample is considered negligible (Moore et al., 2010). The other assumption is a uniform amino-acid composition of sample proteins. Many reports had used 6.25 as a conversion factor based on the 16% average percentage of N in the twenty amino acids by mass (Crook and Casady, 1974). However, these assumptions are violated while quantifying grain samples (Mariotti et al., 2008). For instance, seven percent of the total N in cereals has been reported as non-protein (Fujihara et al., 2008). Moreover, the 16% N is not constant and may change depending on the amino acid composition of the proteins. Different authors have attempted to determine the conversion factor for various foodstuffs. For example, for sorghum, different values were reported as specific conversion factors: 5.93 (Sosulski and Imafidon, 1990), 5.65 (Mosse, 1990), 5.61 (Fujihara et al., 2008), and 6.25 ((Jones and others, 1941) cited in Mariotti et al., 2008). The uncertainty in the conversion factor and, as a result, the lack of precision in estimating true grain protein content has adverse health and economic implications (Moore et al., 2010). Moreover, it adds a challenge dissecting genetic factors controlling grain protein content.

Spectroscopic techniques

Several types of spectroscopic techniques have been utilized to evaluate the total protein content and the protein fractions of grain crops. Grain protein content measurement using this technique is usually employed to obtain a rough estimate of protein content in a purified sample. However, variation in the composition of aromatic amino acids among proteins and the presence of other components that may absorb UV light in the same range as tryptophan and tyrosine biases measurements. Since grain samples contain complex mixtures of proteins and other compounds, it is not a standard method of estimation.

Several colorimetric approaches for quantifying protein have been developed based on the reaction of cupric cations with peptide bonds. The Biuret, Lowry, and BCA methods employ this approach. The Biuret method is a one-step method involving the reduction of cupric (Cu^{2+}) ions to cuprous (Cu^{+}) by peptide bonds present in the protein. This reduction of copper results in the characteristic color, which has an absorption peak at 540 nm. A rapid Biuret method for grain protein content in grain samples has been utilized in crops (Itzhaki and Gill, 1964; Johnson and

Craney, 1971). Lowry and (bicinchoninic acid) BCA methods have increased sensitivity due to the addition of Folin Cicoaltteau reagent and bicinchoninic acid, respectively. Folin Cicoalteau is an antioxidant assay and is mainly used for the estimation of phenolic compounds, including tyrosine and tryptophan. Both the Folin Cicoalteau reagent and bicinchoninic acid form complexes with the reduced Cu^{+} ions. The Folin Cicoalteau and BCA complexes have absorption peaks at 725 to 765 nm and 562 nm (Brenner and Harris, 1995). The incorporation of Folin Cicoalteau and BCA increases Lowry and BCA's sensitivity more than 100-fold, making them ideal for estimating proteins at the micro-level (Afify et al., 2012). BCA has been mainly used in many micro-level estimation protein contents. For example, in sorghum, both Lowry and BCA were employed in the microlevel estimation of soluble proteins (Afify et al., 2012; Sullivan et al., 2018). In maize, BCA was utilized to estimate the protein-level gradient across vitreous and non-vitreous layers (Gayral et al., 2016). The Bradford method is another colorimetric method that uses Coomassie Brilliant Blue to form a protein-dye complex with the advantage of a sample requirement that is 10 to 100 less than BCA, Lowry, and Biuret assays (Chutipongtanate et al., 2012). The assay was used for protein content determination (Steiner et al., 2012). Due to its sensitivity, it was also utilized for assaying protein fraction samples in cereal grains (Geisslitz et al., 2019).

A limitation of colorimetric methods is that proteins to be analyzed must first be extracted. However, other colorimetric methods have been used directly on grain materials (Chan and Wasserman, 1993). The acid orange-dye method is importantly studied mainly in milk but also in grain crops like wheat and soybean (Hymowitz et al., 1969). The procedure used is to bind the acid-orange dye with proteins in the flour. The excess of the dye in washed solution is inversely related to the protein content of the target material (McDonald, 1977). The direct estimation minimizes the workload required for the assay especially in screening numerous samples.

Near-infrared spectroscopy (NIRS) is a non-destructive spectroscopic method routinely used for protein quantification. NIRS measurement is based on the absorption of C-H, O-H, and S-H bonds (Sandorfy et al., 2007). The absorbance at a wide range of spectral frequency (700 nm to 2500 nm) is calibrated using chemometrics methods with a known set of reference samples. For cereal grains and other crops, NIRS can measure grain protein content with higher accuracy and is routinely used (Figueiredo et al., 2006; Alander et al., 2013; Peiris et al., 2019, 2020). However, confounding factors that can influence NIRS results, such as pericarp thickness, moisture, grain

weathering, etc., need to be considered, and calibrations developed accordingly (Guindo et al., 2016). Continued effort to optimize the method to add accuracies, such as the use of large representative samples for calibration, including confounding factors in calibration development and sample sets and proper maintenance of calibrations led to improved accuracy. The fact that NIRS is non-destructive and has a very high throughput compared to other methods makes it very attractive for use in screening large samples.

Factors affecting grain protein content

Environmental factors

The concentration of protein in the grain is the function of three significant elements: N, water, and energy. Both N and energy-rich hydrocarbons are critical for the structural components of amino acids. Moreover, the nitrate uptake and protein synthesis are energy-intensive (Bloom et al., 1992; Cao et al., 2022). The N in plant tissues is a function of N uptake from soil. Apart from N, sulfur is also an essential component of methionine and cysteine, two of the 20 amino acids. Environmental factors such as sulfur fertilization or sulfur availability also dictate grain protein content.

Soil type is one of the environmental factors which determine the availability of soil N. There are two crucial plant nitrogen sources, nitrate and ammonia, and the property of the soil affects their availability and uptake. Nitrate is vulnerable to leaching and surface runoff (Wang and Li, 2019), while ammonia is liable to N loss through denitrification and volatilization (Wang and Li, 2019). Under anaerobic conditions, nitrate also becomes prone to denitrification (Zhu et al., 2013). Moreover, soil texture and pH determine the forms of nitrogen sources and their availability. Coarser, well-aerated soils usually favor nitrification, where the immobile ammonia is oxidized to highly mobile nitrate (Gasser, 1964; Wang and Li, 2019). Ammonium thrives better in clay soils as the high cation exchange capacity and its buffering capacity would limit volatilization. N-recovery directly impacts grain protein, grain yield, and profitability as farmers need to adjust the fertilization to achieve their yield target.

Heat and moisture stress are common environmental factors which commonly coincide with the grain filling stage. Heat stress was associated with increased grain protein content in rice

(Zhen et al., 2019), barley (Ni et al., 2020), and maize (Yang et al., 2018). In maize, post-silking heat stress resulted in reduced activity of enzymes involved in starch synthesis leading to reduced accumulation of starch in the grain (Singletary et al., 1994) and increased grain protein content (Yang et al., 2018). While heat stress does not favor protein synthesis, reduced starch accumulation under heat stress resulted in a higher proportion of protein in the grain (Yang et al., 2018). A detailed study in heat-stressed wheat revealed that the elevated grain protein content was due to the diversion of energy and other metabolic resources to the synthesis of heat shock proteins, while the synthesis of storage proteins, similar to starch synthesis, was hampered by the stress (Wang et al., 2018a).

Drought stress is another environmental condition that influences grain protein content. Moisture is important for several aspects of plant growth, including photosynthesis, and for driving root-mediated absorption of important nutrients from the soil, including nitrogen. Drought stress during grain filling has its own consequence on grain protein. In wheat, drought increases grain protein content (Barutcular et al., 2016; Bana et al., 2018), while it had no impact on rice (Yang et al., 2019b). Increased protein under drought is believed to be due to dilution effect (Rharrabti et al., 2001); however, higher protein under drought and drought-induced senescence may also be achieved through increased remobilization of nitrogen (Yang et al., 2019a). In an experiment that consisted of multiple levels of nitrogen, moisture, and temperature, Campbell et al. (1981) reported that temperature had the highest impact grain protein content in wheat. In that study, the highest protein content was observed under high heat, high N, and moisture-stressed conditions imposed at boot stage. Conditions that favored ample N supply, but low yield resulted in high grain protein content, and the treatment combination that promoted high grain yield resulted in the lowest grain protein content.

Fertilizer management

Studies have shown that grain protein content increases with nitrogen application (Johnson et al., 1973; Bulman and Smith, 1993). Nitrogen fertilizer management that allows better uptake and utilization tends to enhance grain protein content. Split nitrogen application is effective in reducing nitrogen loss and meeting yield targets (Yadav et al., 2017). This practice also improved grain protein percentage in wheat, barley, and sorghum (Bulman and Smith, 1993; Bishnoi et al.,

1995; Blandino et al., 2015; Xue et al., 2016). Likewise, sulfur fertilization along with N has been reported to have a synergistic effect on grain protein content (Järvan et al., 2008; Rossini et al., 2018).

Plant factors

Grain structural somponents and compositional attributes

The major kernel structural components, the bran, the endosperm, and the embryo (germ) of cereals, have varying degrees of protein concentrations. The bran consists of the pericarp, seed coat (testa), and nucellar tissue. The endosperm encompasses the aleurone cell layer and the endosperm cells. The embryo contains the next generation of the plant which grows when sprouted. The embryo has the highest concentration of protein (Ma, 1975; Somavat et al., 2017), followed by the endosperm. Taylor and Schussler (1986) reported the protein percentage of the germ, the endosperm, and the pericarp of two sorghum genotypes to be 17.1 to 18.5%, 7.3 to 10.1 %, and 4.5 to 5.9%, respectively. However, since the endosperm is the largest component of the grain (more than 85% of the grain by weight), the germ's overall protein contribution is smaller compared to the endosperm. In maize, the embryo's share of protein ranged from 8.6 to 20.1% and up to 35.1% in *opaque2* mutants (Landry and Moureaux, 1980; Uribelarrea et al., 2007). The germ in Illinois high protein strains had a greater proportion of kernels by weight than the low protein strains ($P < 0.05$) (Uribelarrea et al., 2004). There is evidence that embryo proportion is under the control of genetic factors in wheat (Golan et al., 2015), maize (Zhang et al., 2012), and rice (Lee et al., 2019). As the embryo is attributed to high grain protein content and overall nutritional quality, such genetic factors need to be investigated to shed light on the possible applicability of the parameter in quality improvement through breeding. The endosperm is often divided into a vitreous or corneous part on the outer section of the endosperm and the inner starchy or floury section. The vitreous portion has a higher protein proportion than the starchy section in maize (Zhang and Xu, 2019), durum wheat (Fu et al., 2018), and sorghum (Ioerger et al., 2007).

The grain protein content, like any concentration parameter, is vulnerable to the dilution effect. Any factor which changes the relative concentration of any of the components would influence grain protein content. This dilution factor is reported as one of the primary sources of

negative correlation between grain yield and grain protein content in wheat (Kibite and Evans, 1984). In an explorative study that compared durum wheat varieties released in different eras, a relative grain protein content decline was reported and was attributed to the dilution effect, not due to a decline in nitrogen uptake or nitrogen harvest index (Motzo et al., 2004). In wheat, Triboi et al. (2006), showed the negative correlation between nitrogen and yield exists irrespective of the environmental effect.

Grain nitrogen is highly impacted by soil nitrogen content and the capacity of plants to extract, assimilate, and transport nitrogen to the grain, which altogether affects grain protein content. At the vegetative stage, root uptake is the primary source of nitrogen to the vegetative sink organs: the growing shoot, stem, leaves, and leaf sheaths. For nitrogen to be used as a structural and functional component, it needs to be assimilated into organic compounds, which is termed nitrogen utilization. As the plant develops, some of the nitrogen is recycled from the older tissue to the younger tissues. The developing reproductive tissue begins to become the sink at flowering, while the vegetative sink starts to serve as an emergent nitrogen supplier. Nitrogen remobilization recycles vegetative nitrogen from the shoot to the developing grain. However, remobilization is not the sole supplier of nitrogen for the grain. For crops like sorghum and maize, post-anthesis nitrogen uptake is also responsible for a significant portion of grain nitrogen. The available nitrogen determines grain nitrogen, and the efficiencies of each of the processes would determine grain nitrogen content. Thus, grain nitrogen content is determined by grain nitrogen and the relative proportion of other grain constituents, including carbohydrate content. Below is a description of major plant structures and functions affecting grain nitrogen content.

Root architecture

Plant root architecture is one of the elements affecting nitrogen and moisture uptake. The architecture determines the volume of soil explored for both nitrogen and moisture. The optimum root architecture for higher nitrogen uptake varies depending on the soil properties, nitrogen status, and moisture conditions. In coarse-textured soils where N is liable to loss through leaching, nitrate tends to regress to the lower soil profile. In such soils, deep-rooted genotypes perform better as they can forage nitrogen from the deeper soil profile (Sullivan et al., 2000; Kristensen and Thorup-Kristensen, 2004; Ehdaie et al., 2010). Deep-rooted winter wheat varieties have been shown to

have better nitrate acquisition from a deeper soil layer (Zhou et al., 2008). Genotypes with dense root architecture at the topsoil and growing deep roots at a later stage may reduce nitrate wastage in well-aerated soils (Dunbabin et al., 2003). Moreover, fewer crown roots in a root system grow deeper and seek more nitrate from a deeper layer (Saengwilai et al., 2014). In maize, deeper crown roots were more active in nitrate transport than the other root types (Dechorgnat et al., 2018). In finer-textured clay soils, leaching is not a detrimental factor for nitrate loss. In such soils, nitrogen occurs mainly as ammonium which is adsorbed to the soil particles and does not leach to deeper soil profiles. Hence, the use of deep-rooted varieties may not have an economic or agronomic advantage in these situations. As a result, investing in a deeper root layer is metabolically costly and with minimal marginal return (Lynch, 2013). A detailed study was conducted on nitrogen source preferences in two maize genotypes selected in two environments. The first line, F44, was adapted to nitrate-leaching prone sandy soil in Florida, and the other, B73, was selected in Iowa on rich Mollisol soil. The root architecture of the two genotypes mirrored the soil properties to which the two genotypes were adapted. Genotype B73 had a deeper root architecture dominated by deeper crown roots with larger volume and surface area compared to F44. Moreover, biochemical, and enzymatic evidence suggested that F44 is well suited for nitrate uptake, whereas B73 was more suited for ammonia uptake (Dechorgnat et al., 2018). In maize, dimorphic phenotypes with shallow seminal roots and deep crown roots determined by axial root angle had consistently higher performance in varying soil and moisture conditions (Dathe et al., 2016). Genotypes with such architecture may perform in a wide range of environments. In sorghum, deeper root structure was correlated with the stay green trait, which renders post-anthesis moisture tolerance and increased nitrogen uptake (Sintayehu et al., 2018). In dry areas, at the later stages of the crop, moisture level and nitrogen (nitrate) regress to deeper layers and deeper root architecture has the ability to forage the resources. The stay green trait in sorghum has been associated with positive nitrogen balance during grain filling (Borrell et al., 2001)

Nitrogen uptake

Nitrogen uptake is mediated mainly by nitrate or ammonia transporters (reviewed in (Forde, 2000; Hao et al., 2020; Tsay et al., 2007; Wang et al., 2018). There are multiple nitrate transporters which function as sensers, transporters and regulators of nitrogen transport from the soil to plants and to different part of the plant organs including leaves and grain (Remans et al.,

2006; Maghiaoui et al., 2020). Plants are capable of foraging nitrate at varying soil nitrate conditions. Plants recruit high-affinity nitrate transports to take up N under low nitrate conditions and low-affinity nitrate transport systems for nitrate uptake under high nitrate conditions. In Arabidopsis, most of the genes in the gene family NRT1 are low-affinity nitrate transporters. Low-affinity transporters are involved in nitrate uptake under high nitrate concentration, above 500 µM. NRT1.1 is an exception from the NRT1 family and has a dual function nitrate affinity through phosphorylation of a specific amino acid, 101^{th} AA, to switch it to a high-affinity transporter. Dephosphorylating the amino acid turns the system into a low-affinity transporter. This protein-level regulation enables plants to rapidly respond to varying external nitrate conditions. In addition, NRT1.1 regulates the high-affinity transporter gene NRT2.1 at the transcription stage by repressing its transcription at high external nitrate conditions (Muños et al., 2004). Both NRT1.1 and NRT2.1 are also involved in lateral root development. NRT1.1 has an additional function of auxin transport. This transport is inhibited by higher nitrate conditions leading to the accumulation of auxin in the root. Auxin stimulates lateral root development. NRT1.1 suppresses lateral root development under low nitrate conditions by allowing basipetal auxin transport and avoiding auxin accumulation (Remans et al., 2006; Maghiaoui et al., 2020). This control of root architecture enables the root system to explore and extract nitrate from soil patches rich in nitrate and avoids expensive resource allocation to low nitrate patches of the soil. In grass species, there are at least two copies of the NRT1.1 gene (Plett et al., 2010). In rice, the OsNRT1.1A and OsNRT1.1B had functionally diverged. OsNRT1.1A had been reported to be expressed in the tonoplast, whereas OsNRT1.1B is localized at the plasma membrane, like the Arabidopsis AtNRT1.1. Unlike AtNRT1.1, NRT1.1A is upregulated by ammonia, an important source of nitrogen in rice, and its over-expression resulted in increased yield and shortened maturity date while its mutant had lower yield and longer maturity (Wang et al., 2018b). There are some natural variants which vary for nitrate uptake activity of NRT1. Comparison of natural variants of one of the nitrate transporters NRT1.1B in the rice subspecies *indica* and *japonica* showed that the indica variant has enhanced nitrogen uptake and transport of nitrogen relative to the *indica* and subsequently resulted in higher nitrogen uptake efficiency and nitrogen use efficiency(Hu et al., 2015).

In Arabidopsis, the ammonium transporter AMT1 is directly involved in root ammonia uptake. Within this family, AtAMT1;1, AtAMT1;2, and AtAMT1;3 explain most of the ammonium transport at the high-affinity range - <1 mM (Hao et al., 2020). Another ammonium

transporter family, AMT2, is involved in the further transport of ammonium to the shoot via xylem loading (Giehl et al., 2017). Ammonium is toxic in plants, and there are multiple venues for ammonium build-up in plant cells, including direct ammonium uptake from the soil, downstream reduction of nitrate, and catabolic pathways of different amino acids. As a result, ammonium uptake and downstream storage and assimilation are strictly regulated. At the transcript level, the three primary ammonium transporters' response to ammonium availability is different. AtAMT1;1 is induced by ammonium availability, AtAMT1;3 has slight upregulation, whereas AtAMT1;2 has constitutive expression. AMT1;3 is also positively regulated by sugar availability (Gazzarrini et al., 1999). Most Ammonium transporters are also post-translationally controlled *via* – (de)phosphorylation to swiftly respond to ammonium status. Phosphorylation of AtAMT1;1 because of ammonium accumulation results in the feedback inhibition of ammonium uptake (Lanquar et al., 2009). AtAMT1;3 is also post-translationally regulated by (de-) phosphorylation at multiple A.A. positions based on ammonium and nitrate levels (Wu et al., 2019). In the presence of ammonium, the ammonium transporters are also involved in the inhibition of nitrate-dependent lateral root growth (Kumar et al., 2020).

Nitrogen utilization

Nitrate can be stored in the root vacuoles, translocated to the shoot to be stored in the vacuoles of aerial organs, or assimilated. Vacuolar sequestration in roots of *Brassica napus* had been associated with reduced nitrogen utilization as it reduces nitrate reductase activity by lowering nitrate flux (Han et al., 2016). Vacuolar nitrate needs to be remobilized back to cytosol before it is loaded to the xylem for transport. Efficient regulation of influx and efflux transporters is significant for maintaining the steady state of nitrate and nitrogen use efficiency. Two proton pumps, vacuolar-type H^+-ATPase (V-ATPase), AtCLCa, and vacuolar proton-translocating inorganic pyrophosphatases, are attributed to nitrogen import into the tonoplast (De Angeli et al., 2006; Krebs et al., 2010; Han et al., 2016). Whereas NPF5.11, NPF5.12, and NPF5.16 are vacuolar nitrate efflux transporters. Mutants for these proton pumps showed increased NUE, increased xylem nitrate, and reduced tissue nitrate than their wild-type counterparts (He et al., 2017). Increased nitrate flux and assimilation seem to contribute to larger photosynthetic capacity, which led to higher biomass yield than those genotypes storing their nitrogen pool as inert nitrate in their vacuoles. In rice, a knock-down of tonoplast localized nitrate transporter OsNPF7.2 drastically

reduced growth under high nitrate conditions (10 mM) but not at low nitrate (1 mM) conditions. OsNPF7.2 is expressed during high nitrate conditions and is expressed at the tonoplast membrane of small and large vacuoles (Hu et al., 2016). OsNPF7.2 did not show any tissue nitrate differences suggesting further evidence is needed to qualify it as an efflux or influx transporter (Hu et al., 2016). Tissue nitrate storage capacity becomes essential in field conditions where the demand for nitrate fluctuates depending on the growth stage and the environment. Two rice varieties contrasting nitrogen use differed in their shoot nitrogen storage. The nitrate reductase activity for both types was similar when 10 mM nitrate was supplied. However, when the nitrogen supply was withdrawn, the nitrate reductase activity of the low nitrate storage variety declined to 80%, while the high storage line maintained its activity (Fan et al., 2007).

Inorganic nitrogen assimilation to organic N

The enzyme nitrate reductase is the first step in the nitrate assimilation and has been postulated to be associated with the nitrogen use efficiency (NUE) of plants. The enzyme's importance was debated in the 1980s and 90s when there was an effort to associate nitrate/nitrite reductase activity to biomass production and grain nitrogen. In a study involving seven diverse sorghum genotypes, Traore and Maranville, (1999) found no association between nitrogen use efficiency at biomass level and grain level with nitrate reductase activity. However, they reported a positive (but not statistically significant) correlation between nitrate reductase activity and grain protein concentration. A similar correlation between grain protein yield and nitrate reductase activity was found in wheat (Eilrich and Hageman, 1973). Another study involving maize hybrids reported a correlation between leaf nitrate reductase activity during ear development and grain protein yield (Deckard et al., 1973).

Internal plant nitrogen status regulates the root nitrate reductase activity by controlling the nitrate flux (Oaks et al., 1977). The nitrate reductase activity increased with increasing nitrate flux (Shaner and Boyer, 1976). Nitrate upregulates the transcription of nitrate reductase (Hoff et al., 1992). An essential role of nitrate reductase in improving nitrate use efficiency and grain yield was recently reported in the indica and japonica subspecies. They have naturally occurred variants where the indica NR2 variant resulted in increased reductase activity (Gao et al., 2019). The increased activity of OsNR2 enhanced nitrate uptake by triggering feedforward transcriptional

upregulation of the indica OsNRT1.1B gene. Similarly, the upregulation of the indica OsNRT1.1B resulted in the upregulation of NR2 gene improving the flux of nitrogen (Gao et al., 2019). Near isogenic lines harboring both variants in japonica background showed increased shoot N, panicle N and grain yield (Gao et al., 2019).

Nitrogen remobilization and post flowering nitrogen uptake

In the vegetative tissue, nitrogen is mainly stored as photosynthetic machinery, Rubisco, the pivotal enzyme in the fixation of carbon during photosynthesis. Rubisco accounts for a quarter of leaf nitrogen and more than half of soluble protein (Mae et al., 1983). Under high nitrogen conditions, plants also store nitrogen in their leaf sheaths, stem, and root. Sorghum plants grown in high nitrate conditions accumulated more than 3-fold of nitrate in their leaf sheathes compared to leaf nitrogen (Worland et al., 2017). During post-anthesis development, these vegetative nitrogen sources are remobilized to fill grain. During grain filling, the high nitrogen demand of the grain cannot be satisfied by the nitrogen uptake only. As a result, plants trigger senescence to retrieve and remobilize nitrogen to the reproductive organ. The resulting stress from nitrogen storage triggers senescence (Pommel et al., 2006). During senescence, the photosynthetic machinery is catabolized by different proteolytic enzymes. Enzymes and transcription factors responsible for senescence have been identified. Cys proteases are reported to be prominent proteolytic enzymes involved in the senescence of leaves. One such protease, HvPAP1, was studied in barley. Over expressing this gene accelerated senescence while silencing it delayed post anthesis senescence (Velasco-Arroyo et al., 2016). The upregulation of these proteases under abiotic stresses also shades light on how different stresses induce senescence. Remobilization in major cereals: rice, wheat, and maize, explains more than half of grain nitrogen (Masclaux et al., 2001). Another proteolytic enzyme associated with senescence is SAG12. In Arabidopsis, SAG12 knock-out lines had a lower nitrogen harvest index (James et al., 2018). Senescence-related transcription factor *GPC-B1* has been found to control grain protein content. In wheat, grain protein content-B1 locus encoding for senescence-related transcription factor was found to increase grain protein. Its introgression into wheat lines caused a significant increase in grain protein content (Tabbita et al., 2017). Catabolized amides, ammonium, and other nitrogen sources are recycled through the action of GS1 (Masclaux-Daubresse et al., 2010). In durum wheat, high grain protein content lines showed high GS1 activity (Nigro et al., 2016).

The recycled or *de novo* synthesized amino acids are translocated to the reproductive sink organs through Amino acid permeases. Amino acid permeases play an essential role in the transport of amino acids through phloem loading and later importing them to the seeds. The *Arabidopsis* AtAAP8 is localized at the phloem's plasma membrane and is involved in the phloem loading of amino acids from source leaves. *aap8* mutants showed decreased amino acid loading and reduced seed numbers while maintaining grain protein content. The authors suggested that AAP8 functions as sink development (Santiago and Tegeder, 2016). AtAAP2 works as an amino acid transporter between xylem and phloem (Zhang et al., 2010). In the grain, the Arabidopsis AtAAP1 is involved in amino acid uptake to the embryo, and its dysfunction resulted in lowering protein content (Sanders et al., 2009).

Post-anthesis/flowering nitrogen uptake contributes significantly to the total grain nitrogen content. In sorghum, split application of high and low levels of nitrogen fertilizer showed that grain protein is mainly dependent on post-anthesis nitrogen supply (Worland et al., 2017). In winter wheat, using radiolabeled ^{15}N showed that a third of grain nitrogen originates from post-anthesis nitrogen uptake (Zhou et al., 2018). In maize, more than half of the grain nitrogen comes from post-silking nitrogen uptake (Coque and Gallais, 2007). Even though post-flowering nitrogen uptake is an important trait for grain protein content, its negative correlation with remobilization efficiency resulted in lower nitrogen harvest index and lower grain nitrogen (Pask, 2009). In sorghum, stay green genotypes have larger post-harvest nitrogen uptake than their senescent counterparts. A study involving 17 sorghum genotypes showed that stay-green genotypes had higher post-anthesis nitrogen uptake compared to senescent types. The senescent types had higher nitrogen remobilized during grain filling (Borrell and Hammer, 2000). Stay-green genotypes, as the name implies, have delayed senescence and hence have a lower nitrogen remobilization rate. Post-harvest nitrate uptake and remobilization rate are negatively correlated (Pask, 2009). A study that evaluated maize genotypes for different nitrogen uptake and remobilization-related traits reported antagonistic QTL clusters for nitrogen uptake and nitrogen remobilization (Coque et al., 2008). One of the drawbacks of remobilization is the degradation of the photosynthetic machinery. As a result, under senescence, nitrogen uptake is severely limited or absent because of the absence of essential input photosynthates. One strategy proposed to increase grain nitrogen is. by utilizing both post-harvest nitrogen uptake and remobilized nitrogen is to delay leaf senescence by increasing stem and leaf sheath nitrogen storage capacity, targeting leaf sheath and true stem for

early remobilization, and inducing leaf remobilization at the later stage of grain filling (Ciampitti and Prasad, 2016; Worland et al., 2017). Zhang et al. (2020) reported that such an approach, together with split nitrogen fertilizer application, resulted in simultaneous improvement of protein concentration and grain yield. This would enable leaves to be photosynthetically active and provide inputs for the roots and the grain. Uribelarrea et al. (2007) reported that 100 years of selection for high protein content in maize resulted in the indirect selection of strains with extensive remobilization and higher pre-and post-flowering nitrate uptake capacities. In wheat, the high-affinity NRT2.1 was associated with post-flowering nitrate uptake (Taulemesse et al., 2015).

Approaches for improving grain protein content

Improving a trait requires genetic variation and selection strategy. The source of genetic variation can either be natural or artificial. Breeders utilize genetic variation to make genetic advances. In maize, long-term divergent selection from a single open-pollinated cultivar in 'Burr's White' was conducted. Continued selection with sizable genetic advances for more than 50 generations created a 14% gap between high and low-grain protein content strains (Dudley, 2007). Wild relatives can also be used to introduce variation in the genetic pool. In wheat, the high grain protein content grain protein content-B1 locus linked to grain protein was sourced mainly from a wild relative var.*dicoccoides* (Blanco et al., 2006; Tabbita et al., 2017). Genome editing, discussed in the last section of this chapter, also provided new tool to modify targeted genome sequences to alter its expression towards specific purpose.

Phenotypic selection

Protein content as a selection criterion has improved grain protein content in many crops. The Illinois selection for high grain protein through direct selection for high grain protein content has made it possible to identify strains with higher protein (Uribelarrea et al., 2007). In sorghum, a four-cycle mass selection increased grain protein by 0.5 percentage points, while negative selection reduced 1.06 percentage points (Ross et al., 1985). In rice, selection in a back cross-population for higher grain protein content and yield resulted in an elite genotype high in grain protein without sacrificing yield (Chattopadhyay et al., 2019). Non-destructive NIR methods have facilitated selection for grain protein. In soybean, it was possible to make early generation selection

on F2 seeds using specialized single seed NIR spectroscopy (Lee et al., 2010). There are at least two drawbacks in using protein content as a selection criterion: its negative correlation with yield and the confounding dilution effect. Selecting merely on protein content usually results in lower grain yield (Ross et al., 1985; Gebre-Mariam and Larter, 1996; Iqbal et al., 2007; Oury and Godin, 2007). Consideration to yield data is important to maintain the agronomic threshold while selecting for high protein.

In addition to grain protein content, the total protein yield, which is the amount of protein produced per unit area, may be of interest. It is computed by the product of average protein concentration with grain yield. Grain protein yield is positively correlated with both grain yield and grain protein content (Kumar et al., 2011; Rhodes et al., 2017). An oat experiment aimed to determine the response to selection for protein yield showed that selection for protein yield-maintained protein content while increasing yield. Selection for high grain yield, on the other hand, raised the non-protein component. But selection for grain protein content leads to a decrease in non-protein grain components (McFerson and Frey, 1992). In another study, selection for protein yield slightly reduced grain protein content (Moser and Frey, 1994). One significant drawback of using protein yield is that most of the variation in protein yield is explained by grain yield, and as a result, high protein content genotypes tend to have less protein yield (McFerson and Frey, 1992).

Monaghan et al., (2001) suggested utilizing grain protein deviations (GPD) - residuals from the regression of grain protein content on grain yield, as a selection criterion for concurrent improvement of grain yield and grain protein content. Cultivars with positive residuals would have higher grain protein content than otherwise predicted from the regression. The heritability of GPD involving more than 70 genotypes from different European countries is moderate (0.44). A larger portion of the GPD variation is explained by genetic components (Mosleth et al., 2020). As the environment confounds grain protein content relationships through GXE, multi-environment data is required to reliably establish the relationship (Oury and Godin, 2007). This parameter has been mainly used in bread and durum wheat (Monaghan et al., 2001). In wheat, post-anthesis nitrogen uptake is positively correlated with GPD (Bogard et al., 2010). Genotypes that have a stable performance for both yield and protein content can be selected using GPD. In an experiment that involved eleven wheat cultivars, GPD was able to identify a cultivar that had relatively stable performance across diverse environments (Marinciu et al., 2018). Contrary to the negative

heterosis for grain protein content, high GPD hybrids had higher grain yield for protein content than line cultivars (Thorwarth et al., 2018). Moreover, QTL mapping using GPD identified QTLs independent of Grain Yield. These QTLs were stable across environments and are colocalized with grain protein content (Nigro et al., 2019).

Nitrogen harvest index (NHI) is the ratio between total grain nitrogen and the total nitrogen in the shoot. Fageria (2014) reviewed the NHI and its relationship with different crop attributes. NHI has positive and significant correlations between yield and grain protein content and can be used as a proxy for selecting genotypes with both high grain yield and grain protein content. However, this correlation does not seem to be global. In durum wheat cultivars, the nitrogen harvest index was not correlated with grain protein content (Desai and Bhatia, 1978). The nitrogen harvest index was positively correlated with yield but not correlated with grain protein content (Löffler and Busch, 1982). It is important to understand which physiological conditions control NHI in the population screened. From its definition, NHI is more likely to be impacted by nitrogen remobilization efficiency (Fageria, 2014) and contrasting nitrogen uptake capacity and remobilization, in theory, may reduce the correlation between NHI and grain protein content. Moreover, the nitrogen status of cultivars impacts nitrogen remobilization efficiency, where genotypes with high nitrogen status have reduced nitrogen remobilization efficiency. The nitrogen Harvest index of two isogenic lines for high protein content locus Gpc-B1 was positively correlated with the presence of the high allele and negatively correlated with straw nitrogen (Tabbita et al., 2013). However, the parameter is cumbersome to be practically used to screen large populations.

To generalize, it is imperative that one must have a good understanding of the architecture of genetic factors controlling grain protein content in order to decide on a given selection strategy. Depending on the species and population, the genetic architecture of grain protein content may be different. In the Illinois long-term selection, the continuous genetic advance from over 50 cycles of selection suggests that polygenic architecture with minor effect is the major contributor to grain protein content (Uribelarrea et al., 2007). Phenotypic selection based on the mean performance of progenies was successful in securing genetic gain. There are also major effect genes controlling grain protein content, with one example being the Gpc-B1 locus in wheat which marker-assisted selection for grain protein content had been successful (Tabbita et al., 2017).

Genomic selection

Genomic selection utilizes genome-wide markers to estimate the additive effect of all loci to predict the genomic estimated breeding value. The applicability of genomic selection for grain protein content was evaluated in wheat. Prediction accuracy estimated as correlation, r, between the observed and predicted values, of r=0.769 was obtained using genomic selection for grain protein. The authors also reported that genomic selection using selection indices enables concurrent selection of grain yield with a minimal penalty in grain protein (Michel et al., 2019). For soybeans, the genomic selection model has improved prediction accuracy (Duhnen et al., 2017). Since phenotyping for grain protein is a significant hurdle, especially in large breeding populations, breeding programs would benefit from optimized genomic selection platforms and improved computational capacity. Integrating genomic selection with the NIR system for phenotyping may further revolutionize breeding for protein content and quality.

Genetic modification

With the advent of molecular technology, it is now possible to make targeted genetic modifications. Overexpression under strong promoters, copy number modification, knock out through T-DNA insertion, and knock-down using RNAi technology had been used to bring phenotypic change. The CRISPR/Cas9 technology now enables selectively editing and modifying genes or regulatory sequences. The technology is widely used in many crop species (Zhang et al., 2018b). Moreover, gene editing had been used to improve grain protein content. In wheat, CRISPR/Cas9 mediated editing of the B and D homeologs of locus TaGW2 increased grain protein content while reducing seed weight (Zhang et al., 2018a). The advantage of the transgenic approach is that beneficial genes discovered can be applied across the taxonomic barriers. This would prove valuable as knowledge from model crops would be transferred to other economically important but underfunded crops.

From the point of improving grain protein content, the primary challenge is to identify target genes that fit the pipeline for improving grain protein. As it is shown in this review, grain protein concentration cannot be seen separate from the background physiological processes which regulate it. For example, high vegetative nitrogen uptake and utilization do not correlate well with

high grain nitrogen. As a result, increasing nitrogen uptake capacity without the sink strength to derive higher remobilization may be a futile exercise. Strategies that strengthen both vegetative and, more importantly, reproductive sink seem to promise to increase grain protein potential (Tegeder and Masclaux-Daubresse, 2018). Vegetative sink strength arises from vigorous vegetative growth and nitrogen storage. Robust vegetative growth was shown to derive nitrogen uptake in maize (Tian et al., 2006). Nitrogen uptake under high nitrogen supply is more associated with the vegetative sink (Peng et al., 2010). Increasing vegetative sink strength through nitrate storage in leaves was suggested to have contributed to nitrate use efficiency in maize RIL lines (Hirel et al., 2001; Tegeder and Masclaux-Daubresse, 2018). Attempts to overexpress nitrate transporters through transgenic approach have shown to increase overall nitrate uptake and nitrogen use efficiency. We saw in the previous sections that the overexpression of nitrogen transporters may trigger higher nitrogen uptake utilization efficiencies. The resulting vigor in photosynthetic capacity may result in enhanced biological performance. The reproductive sink can be strengthened through improved nitrate and amino acid transport capacities to the grain and higher grain biosynthesis activities. In maize, a comparison of two genotypes contrasting for nitrogen use efficiency were found to have a contrasting expression of enzyme coding genes involved in amino acid biosynthesis and interconversion. Increased transport of nitrogen to the grain improved the sink strength (Cañas et al., 2009). In legumes, upregulating pea amino acid transporter amino acid permease gene AAP1 allocated more nitrogen to the shoot and then to the seeds resulting in higher grain protein and grain yield. Nitrogen uptake efficiency and nitrogen use efficiency were improved independently of nitrogen supply level (Perchlik and Tegeder, 2017). In *Vicia faba,* the AAP1 gene's ectopic expression resulted in 10 to 25% increment in grain protein content, 20 to 30% increment in seed size while maintaining starch content. Moreover, radio labeling nitrogen supply from root showed higher labeled nitrogen in the seeds implying sink strength derived the increased nitrogen uptake. The investigators suggested that nitrogen transport into seed is rate-limiting stage and determined seed sink strength (Rolletschek et al., 2005). In wheat, the allelic difference in the gene TaAAP6 explained grain protein content differences (Jin et al., 2018). In rice, overexpression of the amino acid transporter *OsAAP6* elevated grain protein content as compared to its near-isogenic line facilitating amino-acids import to seed endosperm (Peng et al., 2014). Using CRISPR/Cas9 system, Wang et al. (2020b) showed that targeted mutagenesis and silencing of *OsAAP6* gene reduced grain protein content in rice. These result

show that genetic modifications can target specific genes to further modify grain protein content as needed.

Conclusion

The major challenge with improving grain protein content is its negative association with grain yield and the lack of incentive for farmers to risk the yield penalty for increasing grain protein content. Physiological mechanisms such as nitrogen remobilization and the relatively higher energy requirement of protein transport and synthesis make the relationship between overall grain yield and grain protein content negative. However, there are opportunities for improving grain protein content while limiting the penalty on grain yield. The role of environment in grain protein content, and thus, breeding strategies should be a critical consideration. Piece-mill approach using different sources of beneficial traits at each stage of nitrogen metabolism, and overall photosynthetic efficiency and fine-tuning the combinations through successive selection can yield elite lines which satisfy both yield and protein targets.

Reference

Afify, A.E.-M.M.M.R.R., H.S. El-Beltagi, S.M. El-Salam, A.A. Omran, S.M. Abd El-Salam, et al. 2012. Protein solubility, digestibility and fractionation after germination of sorghum varieties. PLoS One 7(2): e31154. doi: 10.1371/journal.pone.0031154.

Alander, J.T., V. Bochko, B. Martinkauppi, S. Saranwong, and T. Mantere. 2013. A Review of Optical Nondestructive Visual and Near-Infrared Methods for Food Quality and Safety. Int J Spectrosc 2013: 1–36. doi: 10.1155/2013/341402.

De Angeli, A., D. Monachello, G. Ephritikhine, J.M. Frachisse, S. Thomine, et al. 2006. The nitrate/proton antiporter AtCLCa mediates nitrate accumulation in plant vacuoles. Nature 442(7105): 939–942.

Arendt, E.K., A. Morrissey, M.M. Moore, and F. Dal Bello. 2008. Gluten-free breads. Gluten-free cereal products and beverages. Elsevier. p. 289--VII

Awika, J.M., and L.W. Rooney. 2004. Sorghum phytochemicals and their potential impact on human health. Phytochemistry 65(9): 1199–1221.

Bana, R.S., S. Sepat, K.S. Rana, V. Pooniya, and A.K. Choudhary. 2018. Moisture-stress management under limited and assured irrigation regimes in wheat (Triticum aestivum): Effects on crop productivity, water use efficiency, grain quality, nutrient acquisition and soil fertility. Indian J Agric Sci 86: 1606–1612.

Barutcular, C., M. Yildirim, M. Koc, H. Dizlek, C. Akinci, et al. 2016. Quality traits performance of bread wheat genotypes under drought and heat stress conditions. Fresen. Environ. Bull 25(12a): 6159–6165.

Bekkerman, A. 2021. Quality forecasts: Predicting when and how much markets value higher-protein wheat. Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie 69(4): 465–490.

Bishnoi, U.R., D.A. Mays, and A. Maiga. 1995. Influence of split-applied nitrogen on grain yield and protein content in ten grain sorghum cultivars. J Plant Nutr 18(6): 1081–1086. doi: 10.1080/01904169509364964.

Blanco, A., R. Simeone, and A. Gadaleta. 2006. Detection of QTLs for grain protein content in durum wheat. Theoretical and Applied Genetics 112(7): 1195–1204.

Blandino, M., P. Vaccino, and A. Reyneri. 2015. Late-season nitrogen increases improver common and durum wheat quality. Agron J 107(2): 680–690. doi: 10.2134/agronj14.0405.

Bloom, A.J., S.S. Sukrapanna, and R.L. Warner. 1992. Root respiration associated with ammonium and nitrate absorption and assimilation by barley. Plant Physiol 99(4): 1294–1301.

Bogard, M., V. Allard, M. Brancourt-Hulmel, E. Heumez, J.-M.M. Machet, et al. 2010. Deviation from the grain protein concentration--grain yield negative relationship is highly correlated to post-anthesis N uptake in winter wheat. J Exp Bot 61(15): 4303–4312. doi: 10.1093/jxb/erq238.

Borrell, A.K., and G.L. Hammer. 2000. Nitrogen dynamics and the physiological basis of stay-green in Sorghum. Crop Sci 40(5): 1295–1307. doi: 10.2135/cropsci2000.4051295x.

Borrell, A., G. Hammer, and E. Oosterom. 2001. Stay-green: A consequence of the balance between supply and demand for nitrogen during grain filling? Annals of Applied Biology 138(1): 91–95. doi: 10.1111/j.1744-7348.2001.tb00088.x.

Brenner, A.J., and E.D. Harris. 1995. A quantitative test for copper using bicinchoninic acid. Anal Biochem 226(1): 80–84.

Bulman, P., and D.L. Smith. 1993. Grain Protein Response of Spring Barley to High Rates and Post-Anthesis Application of Fertilizer Nitrogen. Agron J 85(6): 1109–1113. doi: 10.2134/agronj1993.00021962008500060003x.

Cambell, C.A., H.R. Davidson, and G.E. Winkleman. 1981. Effect of nitrogen, temperature, growth stage and duration of moisture stress on yield components and protein content of manitou spring wheat. Canadian Journal of Plant Science 61(3): 549–563. doi: 10.4141/cjps81-078.

Cañas, R.A., I. Quilleré, A. Christ, and B. Hirel. 2009. Nitrogen metabolism in the developing ear of maize (Zea mays): Analysis of two lines contrasting in their mode of nitrogen management. New Phytologist 184(2): 340–352. doi: 10.1111/j.1469-8137.2009.02966.x.

Cao, H., O. Duncan, and A.H. Millar. 2022. The molecular basis of cereal grain proteostasis. Essays Biochem 66(2): 243–253.

Cato, L., and D. Mullan. 2020. Wheat quality: Wheat breeding and quality testing in Australia. Breadmaking. Elsevier. p. 221–259

Chan, K.-Y., and B.P. Wasserman. 1993. Direct colorimetric assay of free thiol groups and disulfide bonds in suspensions of solubilized and particulate cereal proteins. Cereal Chem 70: 22.

Chattopadhyay, K., S. Sharma, T.B. Bagchi, B. Mohanty, S.S. Sardar, et al. 2019. High-protein rice in high-yielding background, cv. Naveen.

Chutipongtanate, S., K. Watcharatanyatip, T. Homvises, K. Jaturongkakul, and V. Thongboonkerd. 2012. Systematic comparisons of various spectrophotometric and colorimetric methods to measure concentrations of protein, peptide and amino acid: detectable limits, linear dynamic ranges, interferences, practicality and unit costs. Talanta 98: 123–129.

Ciampitti, I.A., and P. V Prasad. 2016. Historical synthesis-analysis of changes in grain nitrogen dynamics in sorghum. Front Plant Sci 7: 275.

Coque, M., and A. Gallais. 2007. Genetic variation for nitrogen remobilization and postsilking nitrogen uptake in maize recombinant inbred lines: heritabilities and correlations among traits. Crop Sci 47(5): 1787–1796.

Coque, M., A. Martin, J.B. Veyrieras, B. Hirel, and A. Gallais. 2008. Genetic variation for N-remobilization and postsilking N-uptake in a set of maize recombinant inbred lines. 3. QTL detection and coincidences. Theoretical and Applied Genetics 117(5): 729–747. doi: 10.1007/s00122-008-0815-2.

Crook, Wayne.J., and A.J. Casady. 1974. Heritability and Interrelationships of Grain-Protein Content with Other Agronomic Traits of Sorghum1. Crop Sci 14(5): 622. doi: 10.2135/cropsci1974.0011183X001400050005x.

Dathe, A., J.A. Postma, M.B. Postma-Blaauw, and J.P. Lynch. 2016. Impact of axial root growth angles on nitrogen acquisition in maize depends on environmental conditions. Ann Bot 118(3): 401–414.

Dechorgnat, J., K.L. Francis, K.S. Dhugga, J.A. Rafalski, S.D. Tyerman, et al. 2018. Root Ideotype Influences Nitrogen Transport and Assimilation in Maize. Front Plant Sci 9: 531. doi: 10.3389/fpls.2018.00531.

Deckard, E.L., R.J. Lambert, and R.H. Hageman. 1973. Nitrate Reductase Activity in Corn Leaves as Related to Yields of Grain and Grain Protein 1. Crop Sci 13(3): 343–350.

Desai, R.M., and C.R. Bhatia. 1978. Nitrogen uptake and nitrogen harvest index in durum wheat cultivars varying in their grain protein concentration. Euphytica 27(2): 561–566.

Dexter, J.E., K.R. Preston, D.G. Martin, and E.J. Gander. 1994. The effects of protein content and starch damage on the physical dough properties and bread-making quality of Canadian durum wheat. J Cereal Sci 20(2): 139–151.

Dhaka, V., and B.S. Khatkar. 2015. Effects of gliadin/glutenin and HMW-GS/LMW-GS ratio on dough rheological properties and bread-making potential of wheat varieties. J Food Qual 38(2): 71–82.

Dudley, J.W. 2007. From means to QTL: The Illinois long-term selection experiment as a case study in quantitative genetics. Crop Sci 47: S--20.

Duhnen, A., A. Gras, S. Teyssèdre, M. Romestant, B. Claustres, et al. 2017. Genomic selection for yield and seed protein content in soybean: a study of breeding program data and assessment of prediction accuracy. Crop Sci 57(3): 1325–1337.

Dunbabin, V., A. Diggle, and Z. Rengel. 2003. Is there an optimal root architecture for nitrate capture in leaching environments? Plant Cell Environ 26(6): 835–844. doi: 10.1046/j.1365-3040.2003.01015.x.

Dykha, M. V, V. Kuzina, and K. Serdyukov. 2021. Grain pricing in Ukraine: A case study of malted barley.

Ehdaie, B., D.J. Merhaut, S. Ahmadian, A.C. Hoops, T. Khuong, et al. 2010. Root system size influences water-nutrient uptake and nitrate leaching potential in wheat. J Agron Crop Sci 196(6): 455–466.

Eilrich, G.L. t, and R.H. Hageman. 1973. Nitrate Reductase Activity and its Relationship to Accumulation of Vegetative and Grain Nitrogen in Wheat (Triticum aestivum L.) 1. Crop Sci 13(1): 59–66.

Fageria, N.K. 2014. Nitrogen harvest index and its association with crop yields. J Plant Nutr 37(6): 795–810. doi: 10.1080/01904167.2014.881855.

Fan, X., L. Jia, Y. Li, S.J. Smith, A.J. Miller, et al. 2007. Comparing nitrate storage and remobilization in two rice cultivars that differ in their nitrogen use efficiency. J Exp Bot 58(7): 1729–1740.

FAOSTAT. 2020. FAOSTAT. FAOSTAT database. http://www.fao.org/faostat/en/#data (accessed 31 July 2020).

Fenster, C. 2003. White food sorghum in the American diet. US Grains Council 43rd Board of Delegates' Meeting

Figueiredo, L.F. de A., F. Davrieux, G. Fliedel, J.F. Rami, J. Chantereau, et al. 2006. Development of NIRS Equations for Food Grain Quality Traits through Exploitation of a Core Collection of Cultivated Sorghum. Journal of Agri 54(22): 8501–8509. doi: 10.1021/JF061054G.

Forde, B. G. (2000). Nitrate transporters in plants: Structure, function and regulation. In *Biochimica et Biophysica Acta - Biomembranes* (Vol. 1465, Issues 1–2, pp. 219–235). Elsevier. https://doi.org/10.1016/S0005-2736(00)00140-1

Fu, B.X., K. Wang, B. Dupuis, D. Taylor, and S. Nam. 2018. Kernel vitreousness and protein content: Relationship, interaction and synergistic effects on durum wheat quality. J Cereal Sci 79: 210–217.

Fujihara, S., H. Sasaki, Y. Aoyagi, and T. Sugahara. 2008. Nitrogen-to-protein conversion factors for some cereal products in Japan. J Food Sci 73(3): C20–C209.

Gao, Z., Y. Wang, G. Chen, A. Zhang, S. Yang, et al. 2019. The indica nitrate reductase gene OsNR2 allele enhances rice yield potential and nitrogen use efficiency. Nat Commun 10(1): 1–10. doi: 10.1038/s41467-019-13110-8.

Gasser, J.K.R. 1964. Some factors affecting losses of ammonia from urea and ammonium sulphate applied to soils. Journal of Soil Science 15(2): 258–272.

Gayral, M., C. Gaillard, B. Bakan, M. Dalgalarrondo, K. Elmorjani, et al. 2016. Transition from vitreous to floury endosperm in maize (Zea mays L.) kernels is related to protein and starch gradients. J Cereal Sci 68: 148–154.

Gazzarrini, S., L. Lejay, A. Gojon, O. Ninnemann, W.B. Frommer, et al. 1999. Three functional transporters for constitutive, diurnally regulated, and starvation-induced uptake of ammonium into Arabidopsis roots. Plant Cell 11(5): 937–947. doi: 10.1105/tpc.11.5.937.

Gebre-Mariam, H., and E.N. Larter. 1996. Genetic response to index selection for grain yield, kernel weight and per cent protein in four wheat crosses. Plant breeding 115(6): 459–464.

Geisslitz, S., C.F.H. Longin, K.A. Scherf, and P. Koehler. 2019. Comparative study on gluten protein composition of ancient (einkorn, emmer and spelt) and modern wheat species (durum and common wheat). Foods 8(9): 409.

Giehl, R.F.H., A.M. Laginha, F. Duan, D. Rentsch, L. Yuan, et al. 2017. A Critical Role of AMT2;1 in Root-To-Shoot Translocation of Ammonium in Arabidopsis. Mol Plant 10: 1449–1460. doi: 10.1016/j.molp.2017.10.001.

Golan, G., A. Oksenberg, and Z. Peleg. 2015. Genetic evidence for differential selection of grain and embryo weight during wheat evolution under domestication. J Exp Bot 66(19): 5703–5711.

Guindo, D., F. Davrieux, N. Teme, M. Vaksmann, M. Doumbia, et al. 2016. Pericarp thickness of sorghum whole grain is accurately predicted by NIRS and can affect the prediction of other grain quality parameters. J Cereal Sci 69: 218–227. doi: 10.1016/J.JCS.2016.03.008.

Haileselassie, M., G. Redae, G. Berhe, C.J. Henry, M.T. Nickerson, et al. 2020. Why are animal source foods rarely consumed by 6-23 months old children in rural communities of Northern Ethiopia? A qualitative study. PLoS One 15(1): e0225707.

Han, Y.-L., H.-X. Song, Q. Liao, Y. Yu, S.-F. Jian, et al. 2016. Nitrogen use efficiency is mediated by vacuolar nitrate sequestration capacity in roots of Brassica napus. Plant Physiol 170(3): 1684–1698.

Hao, D.-L.L., J.-Y.Y. Zhou, S.-Y.Y. Yang, W. Qi, K.-J.J. Yang, et al. 2020. Function and Regulation of Ammonium Transporters in Plants. Int J Mol Sci 21(10): 3557. doi: 10.3390/ijms21103557.

He, Y.-N., J.-S. Peng, Y. Cai, D.-F. Liu, Y. Guan, et al. 2017. Tonoplast-localized nitrate uptake transporters involved in vacuolar nitrate efflux and reallocation in Arabidopsis. Sci Rep 7(1): 1–9.

Henchion, M., M. Hayes, A.M. Mullen, M. Fenelon, and B. Tiwari. 2017. Future protein supply and demand: strategies and factors influencing a sustainable equilibrium. Foods 6(7): 53.

Hirel, B., P. Bertin, I. Quilleré, W. Bourdoncle, C. Attagnant, et al. 2001. Towards a better understanding of the genetic and physiological basis for nitrogen use efficiency in maize. Plant Physiol 125(3): 1258–1270. doi: 10.1104/pp.125.3.1258.

Hoff, T., B.M. Stummann, and K.W. Henningsen. 1992. Structure, function and regulation of nitrate reductase in higher plants. Physiol Plant 84(4): 616–624.

Hu, R., D. Qiu, Y. Chen, A.J. Miller, X. Fan, et al. 2016. Knock-Down of a Tonoplast Localized Low-Affinity Nitrate Transporter OsNPF7.2 Affects Rice Growth under High Nitrate Supply. Front Plant Sci 7(OCTOBER2016): 1529. doi: 10.3389/fpls.2016.01529.

Hu, B., Wang, W., Ou, S., Tang, J., Li, H., Che, R., Zhang, Z., Chai, X., Wang, H., Wang, Y., Liang, C., Liu, L., Piao, Z., Deng, Q., Deng, K., Xu, C., Liang, Y., Zhang, L., Li, L., & Chu, C. (2015). Variation in NRT1.1B contributes to nitrate-use divergence between rice subspecies. *Nature Genetics*, *47*(7), 834–838. https://doi.org/10.1038/ng.3337

Hymowitz, T., F.I. Collins, and S.J. Gibbons. 1969. A Modified Dye-Binding Method for Estimating Soybean Protein 1. Agron J 61(4): 601–603.

Ioerger, B., S.R. Bean, M.R. Tuinstra, J.F. Pedersen, J. Erpelding, et al. 2007. Characterization of polymeric proteins from vitreous and floury sorghum endosperm. J Agric Food Chem 55(25): 10232–10239.

Iqbal, M., A. Navabi, D.F. Salmon, R.-C. Yang, and D. Spaner. 2007. Simultaneous selection for early maturity, increased grain yield and elevated grain protein content in spring wheat. Plant Breeding 126(3): 244–250.

Itzhaki, R.F., and D.M. Gill. 1964. A micro-biuret method for estimating proteins. Anal Biochem 9(4): 401–410. doi: 10.1016/0003-2697(64)90200-3.

James, M., M. Poret, C. Masclaux-Daubresse, A. Marmagne, L. Coquet, et al. 2018. SAG12, a major cysteine protease involved in nitrogen allocation during senescence for seed production in Arabidopsis thaliana. Plant Cell Physiol 59(10): 2052–2063.

Järvan, M., L. Edesi, A. Adamson, L. Lukme, and A. Akk. 2008. The effect of sulphur fertilization on yield, quality of protein and baking properties of winter wheat. Agronomy research 6(2): 459–469.

Jin, X., B. Feng, Z. Xu, X. Fan, Q. Liu, et al. 2018. TaAAP6-3B, a regulator of grain protein content selected during wheat improvement. BMC Plant Biol 18(1): 71.

Johnson, R.M., and C.E. Craney. 1971. Rapid biuret method for protein content in grains. Cereal Chem 48(3): 276–282.

Johnson, V.A., A.F. Dreier, and P.H. Grabouski. 1973. Yield and Protein Responses to Nitrogen Fertilizer of Two Winter Wheat Varieties Differing in Inherent Protein Content of Their Grain 1. Agron J 65(2): 259–263.

Jones, D.B., and others. 1941. Factors for converting percentages of nitrogen in foods and feeds into percentages of proteins.

Kibite, S., and L.E. Evans. 1984. Causes of negative correlations between grain yield and grain protein concentration in common wheat. Euphytica 33(3): 801–810.

Krebs, M., D. Beyhl, E. Görlich, K.A.S. Al-Rasheid, I. Marten, et al. 2010. Arabidopsis V-ATPase activity at the tonoplast is required for efficient nutrient storage but not for sodium accumulation. Proceedings of the National Academy of Sciences 107(7): 3251–3256.

Kristensen, H.L., and K. Thorup-Kristensen. 2004. Uptake of 15 N labeled nitrate by root systems of sweet corn, carrot and white cabbage from 0.2--2.5 meters depth. Plant Soil 265(1–2): 93–100.

Kumar, J., V. Jaiswal, A. Kumar, N. Kumar, R.R. Mir, et al. 2011. Introgression of a major gene for high grain protein content in some Indian bread wheat cultivars. Field Crops Res 123(3): 226–233.

Kumar, V., S.H. Kim, R.A. Priatama, J.H. Jeong, M.R. Adnan, et al. 2020. NH4+ Suppresses NO3−-Dependent Lateral Root Growth and Alters Gene Expression and Gravity Response in OsAMT1 RNAi Mutants of Rice (Oryza sativa). Journal of Plant Biology: 1–17. doi: 10.1007/s12374-020-09263-5.

Landry, J., and T. Moureaux. 1980. Distribution and amino acid composition of protein groups located in different histological parts of maize grain. J Agric Food Chem 28(6): 1186–1191.

Lanquar, V., D. Loqué, F. Hörmann, L. Yuan, A. Bohner, et al. 2009. Feedback inhibition of ammonium uptake by a phospho-dependent allosteric mechanism in Arabidopsis. Plant Cell 21(11): 3610–3622.

Lee, G., R. Piao, Y. Lee, B. Kim, J. Seo, et al. 2019. Identification and characterization of LARGE EMBRYO, a new gene controlling embryo size in rice (Oryza sativa L.). Rice 12(1): 1–12.

Lee, J.-D., J.G. Shannon, and M.-G. Choung. 2010. Selection for protein content in soybean from single F 2 seed by near infrared reflectance spectroscopy. Euphytica 172(1): 117–123.

Li, Y., J. Fu, Q. Shen, and D. Yang. 2020. High-molecular-weight glutenin subunits: Genetics, structures, and relation to end use qualities. Int J Mol Sci 22(1): 184.

Löffler, C.M., and R.H. Busch. 1982. Selection for grain protein, grain yield, and nitrogen partitioning efficiency in hard red spring wheat 1. Crop Sci 22(3): 591–595.

Lynch, J.P. 2013. Steep, cheap and deep: an ideotype to optimize water and N acquisition by maize root systems. Ann Bot 112(2): 347–357.

Ma, E.K.-C. 1975. Morphological and anatomical development of sorghum seed.

Mae, T., A. Makino, and K. Ohira. 1983. Changes in the amounts of ribulose bisphosphate carboxylase synthesized and degraded during the life span of rice leaf (Oryza sativa L.). Plant Cell Physiol 24(6): 1079–1086.

Maghiaoui, A., E. Bouguyon, C. Cuesta, F. Perrine-Walker, C. Alcon, et al. 2020. The Arabidopsis NRT1. 1 transceptor coordinately controls auxin biosynthesis and transport to regulate root branching in response to nitrate. J Exp Bot.

Marinciu, C.M., G. Serban, G. Ittu, P. Mustuațea, V. Mandea, et al. 2018. A new gene source for high positive deviations of grain protein concentration from the regression on yield in winter wheat. Rom Agric Res 35: 71–80.

Mariotti, F., D. Tomé, and P.P. Mirand. 2008. Converting nitrogen into protein—beyond 6.25 and Jones' factors. Crit Rev Food Sci Nutr 48(2): 177–184.

Masclaux, C., I. Quillere, A. Gallais, and B. Hirel. 2001. The challenge of remobilisation in plant nitrogen economy. A survey of physio-agronomic and molecular approaches. Annals of Applied Biology 138(1): 69–81.

Masclaux-Daubresse, C., F. Daniel-Vedele, J. Dechorgnat, F. Chardon, L. Gaufichon, et al. 2010. Nitrogen uptake, assimilation and remobilization in plants: challenges for sustainable and productive agriculture. Ann Bot 105(7): 1141–1157.

McDonald, C.E. 1977. Methods of Protein Analysis and Variation in Protein Results. Farm Research; 34: 5; May/Jun 1977.

McFerson, J.K., and K.J. Frey. 1992. Correlated response to selection for protein yield in oats after three cycles of recurrent selection. Plant breeding 108(2): 149–161.

Michel, S., F. Löschenberger, C. Ametz, B. Pachler, E. Sparry, et al. 2019. Combining grain yield, protein content and protein quality by multi-trait genomic selection in bread wheat. Theoretical and Applied Genetics 132(10): 2767–2780.

Miller, L., and J.A. Houghton. 1945. Micro-kjeldahl determination nitrogen content of amino acids and proteins. Journal of Biological Chemistry 142(1): 374–383.

Monaghan, J.M., J.W. Snape, A.J.S. Chojecki, and P.S. Kettlewell. 2001. The use of grain protein deviation for identifying wheat cultivars with high grain protein concentration and yield. Euphytica 122(2): 309–317. doi: 10.1023/A:1012961703208.

Moore, J.C., J.W. DeVries, M. Lipp, J.C. Griffiths, and D.R. Abernethy. 2010. Total protein methods and their potential utility to reduce the risk of food protein adulteration. Compr Rev Food Sci Food Saf 9(4): 330–357.

Moser, H.S., and K.J. Frey. 1994. Direct and correlated responses to three S 1-recurrent selection strategies for increasing protein yield in oat. Euphytica 78(1–2): 123–132.

Mosleth, E.F., M. Lillehammer, T.K. Pellny, A.J. Wood, A.B. Riche, et al. 2020. Genetic variation and heritability of grain protein deviation in European wheat genotypes. Field Crops Res 255: 107896.

Mosse, J. 1990. Nitrogen-to-protein conversion factor for ten cereals and six legumes or oilseeds. A reappraisal of its definition and determination. Variation according to species and to seed protein content. J Agric Food Chem 38(1): 18–24.

Motzo, R., S. Fois, and F. Giunta. 2004. Relationship between grain yield and quality of durum wheats from different eras of breeding. Euphytica 140(3): 147–154.

Muños, S., C. Cazettes, C. Fizames, F. Gaymard, P. Tillard, et al. 2004. Transcript profiling in the chl1-5 mutant of Arabidopsis reveals a role of the nitrate transporter NRT1. 1 in the regulation of another nitrate transporter, NRT2. 1. Plant Cell 16(9): 2433–2447. doi: 10.1105/tpc.104.024380.

Ni, S.-J., H.-F. Zhao, and G.-P. Zhang. 2020. Effects of post-heading high temperature on some quality traits of malt barley. J Integr Agric 19(11): 2674–2679.

Nigro, D., S. Fortunato, S.L. Giove, A. Paradiso, Y.Q. Gu, et al. 2016. Glutamine synthetase in durum wheat: genotypic variation and relationship with grain protein content. Front Plant Sci 7: 971.

Nigro, D., A. Gadaleta, G. Mangini, P. Colasuonno, I. Marcotuli, et al. 2019. Candidate genes and genome-wide association study of grain protein content and protein deviation in durum wheat. Planta 249(4): 1157–1175.

Oaks, A., M. Aslam, and I. Boesel. 1977. Ammonium and amino acids as regulators of nitrate reductase in corn roots. Plant Physiol 59(3): 391–394.

Oury, F.-X., and C. Godin. 2007. Yield and grain protein concentration in bread wheat: how to use the negative relationship between the two characters to identify favourable genotypes? Euphytica 157(1–2): 45–57.

Papageorgiou, M., and A. Skendi. 2018. Introduction to cereal processing and by-products. Sustainable Recovery and Reutilization of Cereal Processing By-Products. Elsevier. p. 1–25

Pask, A. 2009. Optimising nitrogen storage in wheat canopies for genetic reduction in fertiliser nitrogen inputs.

Peiris, K.H.S., S.R. Bean, A. Chiluwal, R. Perumal, and S.V.K. Jagadish. 2019. Moisture effects on robustness of sorghum grain protein near-infrared spectroscopy calibration. Cereal Chem 96(4): 678–688. doi: 10.1002/cche.10164.

Peiris, K.H.S., S.R. Bean, and S.V.K. Jagadish. 2020. Extended multiplicative signal correction to improve prediction accuracy of protein content in weathered sorghum grain samples. Cereal Chem 97(5): 1066–1074.

Peng, B., H. Kong, Y. Li, L. Wang, M. Zhong, et al. 2014. OsAAP6 functions as an important regulator of grain protein content and nutritional quality in rice. Nat Commun 5(1): 1–12. doi: 10.1038/ncomms5847.

Peng, Y., J. Niu, Z. Peng, F. Zhang, and C. Li. 2010. Shoot growth potential drives N uptake in maize plants and correlates with root growth in the soil. Field Crops Res 115(1): 85–93. doi: 10.1016/j.fcr.2009.10.006.

Perchlik, M., and M. Tegeder. 2017. Improving plant nitrogen use efficiency through alteration of amino acid transport processes. Plant Physiol 175(1): 235–247. doi: 10.1104/pp.17.00608.

Plett, D., J. Toubia, T. Garnett, M. Tester, B.N. Kaiser, et al. 2010. Dichotomy in the NRT gene families of dicots and grass species. PLoS One 5(12): e15289.

Pommel, B., A. Gallais, M. Coque, I. Quilleré, B. Hirel, et al. 2006. Carbon and nitrogen allocation and grain filling in three maize hybrids differing in leaf senescence. European Journal of Agronomy 24(3): 203–211.

Remans, T., P. Nacry, M. Pervent, T. Girin, P. Tillard, et al. 2006. A central role for the nitrate transporter NRT2. 1 in the integrated morphological and physiological responses of the root system to nitrogen limitation in Arabidopsis. Plant Physiol 140(3): 909–921.

Rharrabti, Y., S. Elhani, V. Martos-Nunez, and L.F. del Moral. 2001. Protein and lysine content, grain yield, and other technological traits in durum wheat under Mediterranean conditions. J Agric Food Chem 49(8): 3802–3807.

Rhodes, D.H., L. Hoffmann, W.L. Rooney, T.J. Herald, S. Bean, et al. 2017. Genetic architecture of kernel composition in global sorghum germplasm. BMC Genomics 18(1): 15. doi: 10.1186/s12864-016-3403-x.

Rolletschek, H., F. Hosein, M. Miranda, U. Heim, K.P. Götz, et al. 2005. Ectopic expression of an amino acid transporter (VfAAP1) in seeds of Vicia narbonensis and pea increases storage proteins. Plant Physiology. American Society of Plant Biologists. p. 1236–1249

Ross, W.M., J.W. Maranville, G.H. Hookstra, and K.D. Kofoid. 1985. Divergent Mass Selection for Grain Protein in Sorghum 1. Crop Sci 25(2): 279–282.

Rossini, F., M.E. Provenzano, F. Sestili, and R. Ruggeri. 2018. Synergistic effect of sulfur and nitrogen in the organic and mineral fertilization of durum wheat: Grain yield and quality traits in the Mediterranean environment. Agronomy 8(9): 189.

Saengwilai, P., X. Tian, and J.P. Lynch. 2014. Low crown root number enhances nitrogen acquisition from low-nitrogen soils in maize. Plant Physiol 166(2): 581–589. doi: 10.1104/pp.113.232603.

Sanders, A., R. Collier, A. Trethewy, G. Gould, R. Sieker, et al. 2009. AAP1 regulates import of amino acids into developing Arabidopsis embryos. The Plant Journal 59(4): 540–552.

Sandorfy, C., R. Buchet, and G. Lachenal. 2007. Principles of molecular vibrations for near-infrared spectroscopy. Near-Infrared Spectroscopy in Food Science and Technology; Ozaki, Y., McClure, WF, Christy, AA, Eds: 11–46.

Santiago, J.P., and M. Tegeder. 2016. Connecting source with sink: the role of Arabidopsis AAP8 in phloem loading of amino acids. Plant Physiol 171(1): 508–521.

Shaner, D.L., and J.S. Boyer. 1976. Nitrate reductase activity in maize (Zea mays L.) leaves: I. Regulation by nitrate flux. Plant Physiol 58(4): 499–504.

Simonne, A.H., E.H. Simonne, R.R. Eitenmiller, H.A. Mills, and C.P.C. III. 1997. Could the Dumas method replace the Kjeldahl digestion for nitrogen and crude protein determinations in foods? J Sci Food Agric 73(1): 39–45.

Singletary, G., R. Banisadr, and P. Keeling. 1994. Heat Stress During Grain Filling in Maize: Effects on Carbohydrate Storage and Metabolism. Functional Plant Biology 21(6): 829. doi: 10.1071/PP9940829.

Sintayehu, S., A. Adugna, M. Fetene, A. Tirfessa, and K. Ayalew. 2018. Study of growth and physiological characters in stay-green QTL introgression Sorghum bicolor (L.) lines under post-flowering drought stress. Cereal Res Commun 46(1): 54–66.

Sissons, M., S. Cutillo, I. Marcotuli, and A. Gadaleta. 2021. Impact of durum wheat protein content on spaghetti in vitro starch digestion and technological properties. J Cereal Sci 98: 103156.

Somavat, P., Q. Li, D. Kumar, E.G. de Mejia, W. Liu, et al. 2017. A new lab scale corn dry milling protocol generating commercial sized flaking grits for quick estimation of coproduct yield and composition. Ind Crops Prod 109: 92–100.

Sosulski, F.W., and G.I. Imafidon. 1990. Amino Acid Composition and Nitrogen-to-Protein Conversion Factors for Animal and Plant Foods. J Agric Food Chem 38(6): 1351–1356. doi: 10.1021/jf00096a011.

Steiner, E., A. Auer, T. Becker, and M. Gastl. 2012. Comparison of beer quality attributes between beers brewed with 100% barley malt and 100% barley raw material. J Sci Food Agric 92(4): 803–813.

Sullivan, W.M., Z. Jiang, and R.J. Hull. 2000. Root morphology and its relationship with nitrate uptake in Kentucky bluegrass. Crop Sci 40(3): 765–772.

Sullivan, A.C., P. Pangloli, and V.P. Dia. 2018. Impact of ultrasonication on the physicochemical properties of sorghum kafirin and in vitro pepsin-pancreatin digestibility of sorghum gluten-like flour. Food Chem 240: 1121–1130.

Tabbita, F., S. Lewis, J.P. Vouilloz, M.A. Ortega, M. Kade, et al. 2013. Effects of the G pc-B 1 locus on high grain protein content introgressed into A rgentinean wheat germplasm. Plant Breeding 132(1): 48–52.

Tabbita, F., S. Pearce, and A.J. Barneix. 2017. Breeding for increased grain protein and micronutrient content in wheat: Ten years of the GPC-B1 gene. J Cereal Sci 73: 183–191.

Taulemesse, F., J. Le Gouis, D. Gouache, Y. Gibon, and V. Allard. 2015. Post-flowering nitrate uptake in wheat is controlled by N status at flowering, with a putative major role of root nitrate transporter NRT2. 1. PLoS One 10(3): e0120291.

Taylor, J., J.O. Anyango, and J.R.N. Taylor. 2013. Developments in the science of zein, kafirin, and gluten protein bioplastic materials. Cereal Chem 90(4): 344–357.

Taylor, J.R.N., and L. Schüssler. 1986. The protein compositions of the different anatomical parts of sorghum grain. J Cereal Sci 4(4): 361–369.

Tegeder, M., and C. Masclaux-Daubresse. 2018. Source and sink mechanisms of nitrogen transport and use. New Phytologist 217(1): 35–53.

Thorwarth, P., H.P. Piepho, Y. Zhao, E. Ebmeyer, J. Schacht, et al. 2018. Higher grain yield and higher grain protein deviation underline the potential of hybrid wheat for a sustainable agriculture. Plant Breeding 137(3): 326–337.

Tian, Q., F. Chen, F. Zhang, and G. Mi. 2006. Genotypic Difference in Nitrogen Acquisition Ability in Maize Plants Is Related to the Coordination of Leaf and Root Growth. J Plant Nutr 29(2): 317–330. doi: 10.1080/01904160500476905.

Traore, A., and J.W. Maranville. 1999. Nitrate reductase activity of diverse grain sorghum genotypes and its relationship to nitrogen use efficiency. Agron J 91(5): 863–869.

Triboi, E., P. Martre, C. Girousse, C. Ravel, and A.-M. Triboi-Blondel. 2006. Unravelling environmental and genetic relationships between grain yield and nitrogen concentration for wheat. European Journal of Agronomy 25(2): 108–118.

Tsay, Y.-F., Chiu, C.-C., Tsai, C.-B., Ho, C.-H., & Hsu, P.-K. (2007). Nitrate transporters and peptide transporters. *FEBS Letters*, *581*(12), 2290–2300.Wang, Y.-Y., Cheng, Y.-H., Chen, K.-E., & Tsay, Y.-F. (2018). Nitrate transport, signaling, and use efficiency. *Annual Review of Plant Biology*, *69*, 85–122.

Uribelarrea, M., F.E. Below, and S.P. Moose. 2004. Grain composition and productivity of maize hybrids derived from the Illinois protein strains in response to variable nitrogen supply. Crop Sci 44(5): 1593–1600.

Uribelarrea, M.M., S.P. Moose, and F.E. Below. 2007. Divergent selection for grain protein affects nitrogen use in maize hybrids. Field Crops Res 100(1): 82–90. doi: 10.1016/j.fcr.2006.05.008.

Velasco-Arroyo, B., M. Diaz-Mendoza, J. Gandullo, P. Gonzalez-Melendi, M.E. Santamaria, et al. 2016. HvPap-1 C1A protease actively participates in barley proteolysis mediated by abiotic stresses. J Exp Bot 67(14): 4297–4310.

Verified Market Research. 2019. Wheat Protein Market Size , Share ,Trends |Analysis | Forecast. https://www.verifiedmarketresearch.com/product/wheat-protein-market/ (accessed 18 July 2020).

Wang, X., L. Hou, Y. Lu, B. Wu, X. Gong, et al. 2018a. Metabolic adaptation of wheat grain contributes to a stable filling rate under heat stress. J Exp Bot 69(22): 5531–5545.

Wang, W., B. Hu, D. Yuan, Y. Liu, R. Che, et al. 2018b. Expression of the nitrate transporter gene OsNRT1. 1A/OsNPF6. 3 confers high yield and early maturation in rice. Plant Cell 30(3): 638–651.

Wang, Z.-H., and S.-X. Li. 2019. Nitrate N loss by leaching and surface runoff in agricultural land: A global issue (a review). Advances in agronomy 156: 159–217.

Wang, S., Y. Yang, M. Guo, C. Zhong, C. Yan, et al. 2020. Targeted mutagenesis of amino acid transporter genes for rice quality improvement using the CRISPR/Cas9 system. Crop J 8(3): 457–464.

Worland, B., N. Robinson, D. Jordan, S. Schmidt, and I. Godwin. 2017. Post-anthesis nitrate uptake is critical to yield and grain protein content in Sorghum bicolor. J Plant Physiol 216(February): 118–124. doi: 10.1016/j.jplph.2017.05.026.

Wu, X., T. Liu, Y. Zhang, F. Duan, B. Neuhäuser, et al. 2019. Ammonium and nitrate regulate NH4+ uptake activity of Arabidopsis ammonium transporter AtAMT1; 3 via phosphorylation at multiple C-terminal sites. J Exp Bot 70(18): 4919–4930.

Xue, C., G.S. auf'm Erley, A. Rossmann, R. Schuster, P. Koehler, et al. 2016. Split Nitrogen Application Improves Wheat Baking Quality by Influencing Protein Composition Rather Than Concentration. Front Plant Sci 7(JUNE2016): 738. doi: 10.3389/fpls.2016.00738.

Yadav, M.R., R. Kumar, C.M. Parihar, R.K. Yadav, S.L. Jat, et al. 2017. Strategies for improving nitrogen use efficiency: A review. Agricultural Reviews (OF). doi: 10.18805/ag.v0iof.7306.

Yang, M., M. Geng, P. Shen, X. Chen, Y. Li, et al. 2019a. Effect of post-silking drought stress on the expression profiles of genes involved in carbon and nitrogen metabolism during leaf senescence in maize (Zea mays L.). Plant Physiology and Biochemistry 135: 304–309. doi: 10.1016/j.plaphy.2018.12.025.

Yang, H., X. Gu, M. Ding, W. Lu, and D. Lu. 2018. Heat stress during grain filling affects activities of enzymes involved in grain protein and starch synthesis in waxy maize. Sci Rep 8(1): 1–9.

Yang, X., B. Wang, L. Chen, P. Li, and C. Cao. 2019b. The different influences of drought stress at the flowering stage on rice physiological traits, grain yield, and quality. Sci Rep 9(1): 1–12.

Zhang, P., W.B. Allen, N. Nagasawa, A.S. Ching, E.P. Heppard, et al. 2012. A transposable element insertion within ZmGE2 gene is associated with increase in embryo to endosperm ratio in maize. Theoretical and Applied Genetics 125(7): 1463–1471.

Zhang, Y., D. Li, D. Zhang, X. Zhao, X. Cao, et al. 2018a. Analysis of the functions of Ta GW 2 homoeologs in wheat grain weight and protein content traits. The Plant Journal 94(5): 857–866.

Zhang, L., Z. yuan Liang, X. ming He, Q. feng Meng, Y. Hu, et al. 2020. Improving grain yield and protein concentration of maize (Zea mays L.) simultaneously by appropriate hybrid selection and nitrogen management. Field Crops Res 249: 107754. doi: 10.1016/j.fcr.2020.107754.

Zhang, Y., K. Massel, I.D. Godwin, and C. Gao. 2018b. Applications and potential of genome editing in crop improvement. Genome Biol 19(1): 210.

Zhang, L., Q. Tan, R. Lee, A. Trethewy, Y.-H. Lee, et al. 2010. Altered xylem-phloem transfer of amino acids affects metabolism and leads to increased seed yield and oil content in Arabidopsis. Plant Cell 22(11): 3603–3620.

Zhang, H., and G. Xu. 2019. Physicochemical properties of vitreous and floury endosperm flours in maize. Food Sci Nutr 7(8): 2605–2612.

Zhen, F., W. Wang, H. Wang, J. Zhou, B. Liu, et al. 2019. Effects of short-term heat stress at booting stage on rice-grain quality. Crop Pasture Sci 70(6): 486–498.

Zhou, M.X. 2009. Barley production and consumption. Genetics and improvement of barley malt quality. Springer. p. 1–17

Zhou, B., M.D. Serret, J.B. Pie, S.S. Shah, and Z. Li. 2018. Relative contribution of nitrogen absorption, remobilization, and partitioning to the ear during grain filling in chinese winter wheat. Front Plant Sci 9: 1351.

Zhou, S.L., Y.C. Wu, Z.M. Wang, L.Q. Lu, and R.Z. Wang. 2008. The nitrate leached below maize root zone is available for deep-rooted wheat in winter wheat-summer maize rotation in the North China Plain. Environmental Pollution 152(3): 723–730. doi: 10.1016/j.envpol.2007.06.047.

Zhu, X., M. Burger, T.A. Doane, and W.R. Horwath. 2013. Ammonia oxidation pathways and nitrifier denitrification are significant sources of N2O and NO under low oxygen availability. Proceedings of the National Academy of Sciences 110(16): 6328–6333.

Table 1-1 Grain protein composition of some cereal crops.

Crop	Number	grain protein content Range (%)		Reference
		Min	Max	
Oat	5	9.7	17.3	(Sterna et al., 2016)
	8	13.81	19.29	(Redaelli et al., 2015)
	26	15.3	23.1	(Ahola et al., 2020)
Maize	35	6.67	11.34	(Ünlü et al., 2018)
	92	12.16	14.95	(Vancetovic et al., 2015)
Sorghum	332	4.7	17	(Deosthale et al., 1970)
	390	8.1	18.81	(Rhodes et al., 2017)
	92	3.5	12.6	(Badigannavar et al., 2016)
Durum Wheat	140	12.22	18.11	(Kendal et al., 2019)
Bread Wheat	60	7.39	13.97	(Pronin et al., 2020)
	225	10.43	14.48	(Akcura et al., 2016)
Rice	6	7.44	9.67	(Kaur et al., 2018)
	10 Populations	8.45	9.93	(Eizenga et al., 2014)
Barley	8	10.8	12.3	(Wang et al., 2003)
	158	8.02	13.5	(Cai et al., 2013)

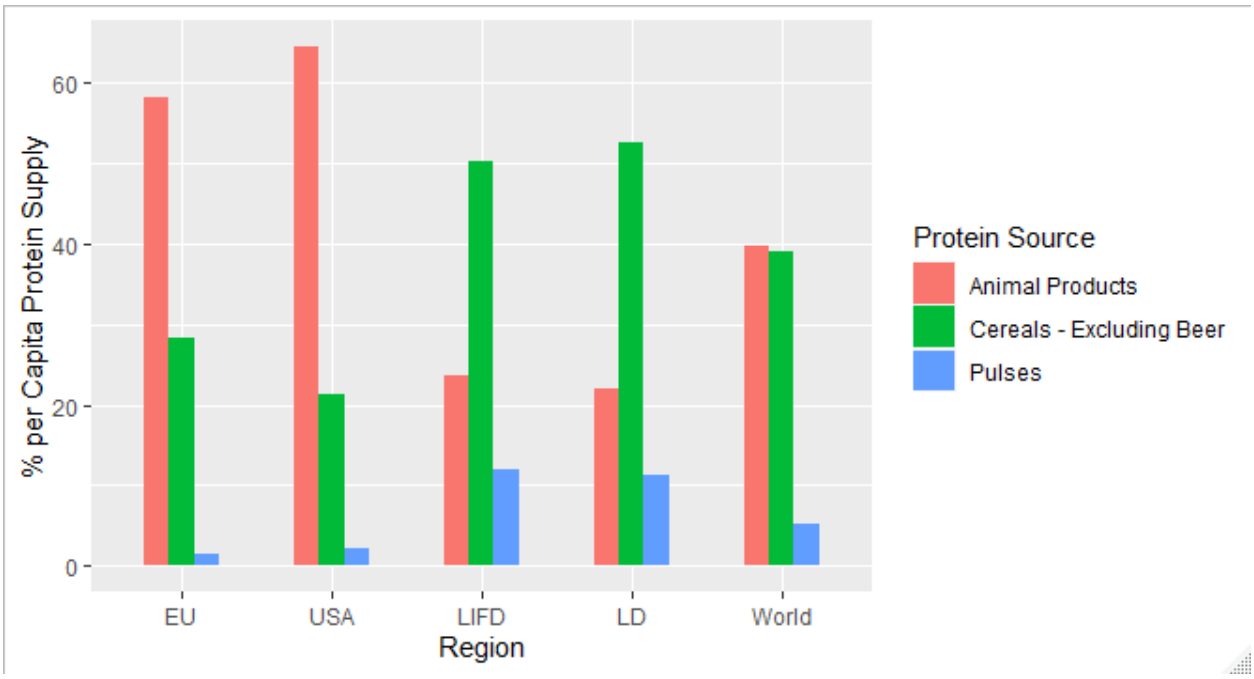


Figure 1-1 Economic status based disaggregated contribution of cereals to the protein supply of the world.

LIFD and LD stand for Low Income Food Deficient Countries and Least Developed Countries, respectively. (Data from FAOSTAT 2020)

Chapter 2 - Adaptation to agroclimatic conditions fashioned some grain physicochemical attributes of sorghum in Ethiopia

Abstract

Sorghum is a significant source of nutrients for people living in drought-affected areas of the world. Unlike other major cereals, sorghum grains develop naked directly exposed to climatic conditions and hence may be vulnerable to climate induced biotic and abiotic stresses. This study was based on the hypothesis that the physicochemical characteristics of sorghum grain may confer fitness to bioclimatic induced stressors. The objective of this study was to assess the association of grain physicochemical parameters and bioclimatic conditions of their collection region. The study utilized Ethiopian sorghum landraces, evaluated them for different grain characteristics and attempted to determine the association between them. Spearman correlation between tannin and various other plant features found negative and significant ($P < 0.001$) correlations including hundred-kernel-weight, virtuousness, and head compactness. Significant associations ($P < 0.01$) were also observed between precipitation gradient and grain attributes. Tannin presence and loose panicle architecture tend to dominate the high precipitation agro-ecologies. This seems to fit the expectation that tannin and looser panicle architecture offer an adaptive advantage in humid areas where grain mold and other fungal pathogens tend to prevail. Grain dimension traits and hundred kernel weight also showed significant correlations with climatic conditions. Membership in any of the five botanical races had an important role in the adaptation and grain characteristics of the landraces. The PCs from the genomic data are also correlated with both climatic variables and grain parameters. Genes in control or directly taking part in the synthesis of polyphenolic compounds had association with the overall adaptation of sorghum landraces. Overall, grain attributes in sorghum have an adaptive role, and careful analysis of target areas is crucial in efforts aimed at improving sorghum grain quality attributes.

Introduction

Ethiopia is located in the North-Eastern Africa region, recognized as the center of diversity of sorghum, where the major cultivated races of the crop have been grown for millennia (Doggett, 1970). Owing to the diverse agroecologies of the country, farmers in Ethiopia had selectively domesticated diverse sets of genotypes suited to their culture, utilization and production practices, and their specific agroecology. Studies on the diversity of Ethiopian sorghum landraces based on morphological attributes (Teshome et al., 1997) and molecular markers (Cuevas and Prom, 2013; Girma et al., 2019; Mengistu et al., 2020) showed the richness of the Ethiopian sorghum gene pool. This pool harbors various beneficial agronomic traits such as high lysine (Singh and Axtell, 1973), high protein content (Rhodes et al., 2017), post-flowering drought tolerance, stay-green (Haussmann et al., 2002), grain mold resistance (Pugh et al., 2017; Nida et al., 2019), sugar cane aphid resistance (Muleta et al., 2021), and resistance to the parasitic weed *Striga* (Abate et al., 2017).

Adaptability to a wide range of environmental conditions is an important factor for the survival of species. The scope of the utilization of a crop is also based on the degree to which it is adapted to a diverse set of agroecology. Under climate change scenarios where weather uncertainty, population pressure, and reduction in per capita land area result in an expansion of crop agriculture to marginal lands, adaptability and ecological plasticity of species are of paramount importance to sustain livelihoods.

Several factors may come into play to influence the adaptability of species and varieties to a given environment. Kernel physico-chemical attributes are particularly important for the adaptation of sorghum. As the genetic material is carried from one generation to the next via a seed, crops have evolved various adaptation mechanisms to maintain this natural cycling of generations. The presence of tannin in sorghum grain is one of the adaptation mechanisms (Morris et al., 2013; Lasky et al., 2015). Tannin has been associated with grain mold resistance in maturing kernels (Melake-Berhan et al., 1996; Nida et al., 2021) and chilling tolerance, especially during germination and emergence (Marla et al., 2019). Moreover, the presence of tannin is often correlated with environmental factors. However, the distribution of tannin sorghums across different agroecology in Ethiopia and its association with other characters is not documented.

Another layer of protection that sorghum kernels have evolved is through alienating pathogens by denying them access to nutrients. The way sorghum protein bodies and starch granules are organized in the endosperm renders sorghum nutrients inaccessible to intruding microbes and insect pests. Even though these characteristics of the crop are detrimental to the overall nutritional quality of sorghum (Duodu et al., 2003), it is an important characteristic for its adaptability and survival. These scenarios need to be addressed in crop improvement approaches. Panicle architecture itself also imparts another barrier to invasion by grain mold pathogens (Sharma et al., 2010) and bird attack (Bullard, 1988). Loose and dropping panicles with extensive glume coverage, such as that of the guinea race, tend to drain moisture quickly and maintain low humidity in the panicle limiting invasion by grain mold pathogens. Thus, it is a crucial adaptation trait in humid regions where panicle and leaf diseases are critical. The sturdy stalk structure of durra sorghums that serves as a sink source during vegetative growth has been hypothesized to serve as a mode of survival under dry conditions through nutrient remobilization during later growth stages (Blum, 2004).

In addition to the morpho-agronomic characteristics, several studies have looked at adaptation signatures at the genome level across the global sorghum collections (Jordan et al., 1979; Morris et al., 2013; Lasky et al., 2015) as well as collections from specific countries (Maina et al., 2018; Faye et al., 2019; Girma et al., 2020). Genomic factors related to maturity (Lasky et al., 2015), tannin (Lasky et al., 2015), inflorescence architecture (Olatoye et al., 2018), drought tolerance (Olatoye et al., 2018; Girma et al., 2020), seed size (Tao et al., 2017; Wang et al., 2020a) and recently sugar cane aphid resistance (Muleta et al., 2021) have been discovered. Ethiopian sorghum landraces provide a unique opportunity for understanding the adaptive role of grain physico-chemical characteristics. The diverse gene pool for the crop, wide-range of agroecology and the long history of cultivation and utilization of the crop may provide a great deal of information about artificial and natural selection pressures and their effects on the genetic and phenotypic variations of Ethiopian landraces across the country. Moreover, larger sample of diverse landraces per local agroecology, which was not possible in the other global studies, may offer better statistical power to understand the impact of local bioclimatic factors in shaping the overall genetic and morphological architecture of the accessions and its signature on adaptation. This study is part of a comprehensive project supported by the USAID FtF program, where over 2000 accessions representing the diverse agroecology of Ethiopia were characterized for various

agronomic traits. The study aims to investigate the relationship between bioclimatic factors and plant attributes and how these may have guided the adaptation and selective cultivation of sorghum in different agroecology of the country.

Material and Methods

Plant materials

The plant materials included in this study are a subset of the Ethiopian sorghum landrace collection representing diverse regions and agroecology of the country. Details of the materials are outlined in (Girma et al., 2019). Briefly, 2010 accessions were drawn out of the over 9000 accessions maintained by the Ethiopian Biodiversity Institute. Released varieties and landraces grown by local farmers in different regions were also included. The materials were planted on a single-row plot at Melkassa Research Center during the 2014 season for seed increase and preliminary observation. For this study, we utilized 1579 materials for which genotype data were available. Of these 1523 were landraces, 10 were elite materials from breeding programs, 32 released cultivars, and seven were introductions. The remaining seven had no collection data.

The materials were grown at multiple locations throughout the country for three subsequent seasons. Due to the size of the entries, the experiment was not replicated per location. The data used for this study were derived from experiments carried out at selected locations representing the ecological diversity of sorghum production, namely Arsi Negele, Haramya, Bako, and Pawe. Data were collected on plant height, days to flowering, and grain mold incidence scored using a 1-5 scale, with one being resistant (no disease) and five for susceptible (Tessema et al., 2019). Panicle compactness was coded following the sorghum descriptor (IBPGR and ICRISAT, 1993) with some modifications as: (1) very lax, very loose (2) loose erect, semi-loose droop, (3) Semi-loose erect (4) semi-compact and compact. The grain samples from the 2016 Arsi-Negele test were evaluated for determining grain characteristics such as thousand kernel weight and endosperm vitreousness following the sorghum descriptors (IBPGR and ICRISAT, 1993). Grain translucence was visually scored under a light box using a 1-5 scale with “1” for completely opaque and “5” for completely translucent. The grain dimension, length (l), width (w) and height (h) were evaluated using a Aickar ™ digital micro-caliper to the 0.01 accuracy from 5 kernels per accession. The presence of

tannin was determined using a bleach test (Waniska et al., 1992). Black coloring after bleaching indicated presence of tannin and was coded as "1", and no black coloring as the absence of tannin "0". Pericarp thickness was evaluated following the procedure outlined in (Gomez et al., 1997) by scraping the kernels. If the pericarp comes out as flakes, it was considered thick and scored "1' and if it fragmented or powdered, it was thin and scored "0".

GBS genotyping

The work employed genotypic data from two sets. One was from the Ethiopian collection, and the other is from the global sorghum collection. For the Ethiopian collection, details of DNA extraction, library preparation, and sequencing were as outlined in (Tessema et al., 2019). The global accessions included in the second set were diverse materials from different sources, and details of the accessions is outlined in (Wang et al., 2020a). The raw fastq sequence for global sorghum collection is a subset of SAMN01828196 BIOSAMPLE from the NCBI Sequence Read Archive database and was obtained using *fastq-dump* from the SRA toolkit (v2.10.8 https://hpc.nih.gov/apps/sratoolkit.html). The raw sequences from the Ethiopian core collection and global accessions were combined for the GBS pipeline. The TASSEL Version 5 (Bradbury et al., 2007), and GBSV2 pipeline (Glaubitz et al., 2014) were used to process the raw sequence files, align to the *Sorghum bicolor* reference genome version 3.1.1 from Phytozome (McCormick et al., 2018) using the "very-sensitive" parameter of Bowtie2 (Langmead and Salzberg, 2012), and called SNPs for the accessions. The GBS pipeline of the Ethiopian core collection and the global sorghum accessions yielded 342,395 SNPs. These were further filtered for bi-allelic sites at a maximum of 20% missing data, and a minimum of 1% minor allele frequency using VCFtools - 0.1.17 (Danecek et al., 2011), finally retaining 311,228 SNPs. Imputation using Beagle4.1 (Browning and Browning, 2016) was performed separately for each chromosome.

Bioclimatic variables

Bioclimatic data was obtained from WorldClim Version2 database (Fick and Hijmans, 2017). Administrative boundaries were retrieved from Open Africa Database (https://africaopendata.org/dataset/ethiopia-shapefiles). In house R code using the R-Package

Raster (Hijmans et al., 2015) was used to extract bioclimatic variables. District level mean value of each variable was used for further analysis.

Correlation and principal component analysis (PCA) among plant and bioclimatic variables

Character association was evaluated using spearman correlation analysis using the function *corr.test* in Psych (v2.0.12; (Revelle and Revelle, 2015) package. Partial correlations were computed using *pcor.test* from *ppcor* package (Kim, 2015). PCA was implemented using the PCA function of FactoMineR (Husson et al., 2016) package. The biplots were displayed using the FactoExtra package (Kassambara and Mundt, 2017).

Population Genomic Analysis

Population structure among the landraces was evaluated using PCA utilizing the PrincipalComponentsPlugin in TASSEL 5.0 (Bradbury et al., 2007). A model-based tool ADMIXTURE (Alexander et al., 2015) was used to infer ancestry estimates using genetic clusters 2 to 20. PLINK (Purcell et al., 2007) using the option –indep-pairwise 50 5 0.5 was used to prune the SNP based on LD, and 65,712 SNPs were retained for estimating ancestry. Population differentiation (F_{st}) among subpopulations was computed using VCFtools - 0.1.17 (Danecek et al., 2011). Nucleotide diversity π was computed for a window of 1 Mb using VCFtools - 0.1.17 (Danecek et al., 2011).

Race membership determination

For the Ethiopian core collection, putative landrace determination was made following two approaches. The major botanical races were manually determined using the spikelet and inflorescence features as outlined in (Harlan and de Wet, 1972). The botanical race assignment of potentially intermediate races and those with ambiguous features were inferred with supervised admixture analysis using ADMIXTURE software (Alexander et al., 2015) utilizing manually assigned landraces and the predetermined botanical races of global collection. An admixture coefficient value of 0.8 was used as a cut-off value for admixture assignment.

Genome-wide association study (GWAS) of plant attributes

Biallelic SNPs with minor allele frequency above 0.05 were utilized for the GWAS of plant attributes. General Linear Model (GLM), Mixed Linear Model (MLM), fixed and random model circulating probability unification (FarmCPU), and Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway (BLINK) using the R package GAPIT (Lipka et al., 2012) were implemented for the association study. False discovery correction rate (FDR) <0.05 was used as a threshold to determine the significance of SNP effect.

Genomic Signatures for local adaptation

A PCA-based statistic was used for identifying outlier SNPs that may be associated with local adaptation. SNPs associated with Principal Components (PCs) which defined population structure were assumed to be potential contenders for local adaptation. A tenth of one percent outlier SNPs were selected as a potential SNPs using loading of the first few PCs analyzed using Tassel (Bradbury et al., 2007).

Priori genes for grain weight, grain quality, and panicle compactness

Priori genes for grain weight and panicle morphology were assembled from (Tao et al., 2017) and (Olatoye et al., 2018), respectively. Moreover, other genes related to anthocyanin regulation and synthesis, kafirin genes were assembled. Linkage with the priori gene was established if the outlier SNP is within 50 kbp flanking region of the priori gene.

Result

Racial attributes of the collections

Botanical race assignment utilized two approaches. The first approach was to manually assign the major botanical races based on kernel features as outlined in (Harlan and de Wet, 1972). Botanical race determination on the remaining unassigned genotypes was made using supervised admixture analysis utilizing the global sorghum diversity collection as a reference (Wang et al., 2020a). Supervised admixture analysis using the software ADMIXTURE has been found reliable

in the literature (Thornton et al., 2014). Admixture, a model-based ancestry estimating software using autosomal SNPs of individuals (Alexander and Lange, 2011), has the option to accept predefined ancestral populations to supervise the learning phase and enhance ancestry estimation of individuals.

Table 2-1 shows the number of genotypes from Ethiopian core population and the global sorghum diversity collection used as a reference for our supervised race assignment.

The result shows that the durra race dominates (48.9%) the Ethiopian collection, followed by the intermediate race durra-bicolor (18.9%) and caudatum (15.4%). Guinea has smaller (0.5 %) representation even though guinea-caudatum mixed race has a significant proportion (13.6%). Kafir was also represented with a smaller percentage (0.8%) in the core collection (Table 2-2).

To further validate the race assignment, we performed PCA, a non-parametric factor reduction method, by forming mutually uncorrelated dimensions which maximize the variances of the first few PCs. The analysis facilitates observation of individual clusters in a space determined by the orthogonal dimensions as an axis. PCA of the accessions showed similar clustering as predicted by the ADMIXTURE analysis (Figure 2-1). The first five PCs explained 22% of the total genetic variation of the Ethiopian accessions and the global collection. Each of these five PCs differed in their importance in discriminating the botanical races. Durra accessions were differentiated from the rest across the PC1 axis while durra-bicolor on PC2. PC3 separated one of the clusters of guinea-caudatum from the rest; PC6 separated the two guinea-caudatum clusters and caudatum clusters from the durra and durra-bicolor cluster

Among the botanical races, the F_{st} showed durra-bicolor as a highly differentiated race relative to other races: bicolor (0.22), durra (0.28), caudatum (0.28), and guinea-caudatum (0.32). The durra and caudatum races were also moderately differentiated (0.22) (Table 2-3).

The level of LD for the whole core collection and across the different botanical races was analyzed. On average, in the Ethiopian collection, LD decayed to half of its maximum r^2 in ~10 kb and to the background LD (r^2 ~ 0.1) by around 46 kb (Figure 2-2). The rate of decay across the botanical races is different. Caudatum decayed faster to its half-maximum r^2 within 12kb and durra

in 13 kb, bicolor (15.1 kb), and durra bicolor (18 kb), while the guinea-caudatum (46 kb) had slower decay.

To evaluate the shift in allele distribution within populations, nucleotide diversity (π) was evaluated for each botanical race separately. A comparison of the median nucleotide diversity, which quantifies the expected per site pair-wise nucleotide differences, showed that the bicolor race had the largest nucleotide diversity, where the median of 50 kb genomic bins had ~6% more nucleotide diversity than the whole population. The rest of the landraces had smaller π than the whole population, where caudatum, durra-bicolor, durra, and guinea-caudatum had 92%, 72%, and 65%, and 50% of the whole core collection.

An optimum number of hypothetical ancestral populations (K) is determined as the K value, which minimizes the cross-validation error (Figure 2-3 A). However, it was difficult to find the optimum K value which satisfies this requirement. Efforts were made to estimate the admixture using K=6, K=8, and K10 (Figure 2-3 B-D). The analysis showed that the Ethiopian core collection is a highly admixed. Moreover, the botanical races themselves are a mixture of subpopulations. For example, Durra at K=6 is a mixture of G1 and G5.

Phenotypic evaluation of grain and panicle traits

Given that Ethiopia is the center of origin and diversity for the crop and is also endowed with wide agroecology (Figure 2-4), sorghum accessions in the country harbor a wide range of variations, including grain and panicle traits (Table 2-4 and Figure 2-5). The grain yield per panicle ranged from no seed set to a high of 222.2 g, with an overall mean of 49.6 g. The mean hundred kernel weight across accessions was 2.5g, with a score ranging from a low of 0.6 g to 4.7 g in bold accessions. The mean grain dimension, length, width, and thickness were 3.9, 2.8, and 2.5 mm, with the range for grain length being 2.6 to 5.1 mm, width 2.3 to 5.6 mm, and thickness 1.7 to 4.7 mm. Grain protein content also ranged from 7.0 to 14.0 %.

Spearman correlation analysis revealed highly significant associations among the plant attributes Table 2-5. Landraces with tannin tend to possess loose panicle ($r = -0.30$, $P < 0.0001$) thicker pericarp ($r = 0.15$, $P < 0.0001$), starchy endosperm ($r = 0.29$, $P < 0.0001$), opaque kernel ($r = 0.34$, $P < 0.0001$), and smaller seed mass ($r = -0.28$, $P < 0.0001$). Large and heavier seeds were

associated with a more compact panicle (r=0.44, P< 0.0001), higher grain protein content (r=0.37, P< 0.0001) and more translucent kernel (r=0.19, P< 0.0001). Compact panicle was also associated with thinner pericarp (r=0.21, P< 0.0001) and translucent kernel (r=0.31, P< 0.0001).

Race membership explained some of the variations among the accessions. The percentage of variance for grain size attributes explained by race and estimated using η^2 revealed that race membership explained a substantial percentage of the variation for the traits. Race had influence on kernel width (28.2 %), HKW (27.5 %), and kernel length (13.1 %). Durras had the largest HKW (2.88 g) while durra-bicolor the smallest (1.88 g). Similarly, kernel dimensions were also variable among the races. Durras had the largest values for kernel length, width, and thickness. Another peculiar observation was the width to the length ratio, which was above 1 for durra while it was below 1 for the rest of races.

A Chi-square test of independence was used to determine the association between categorical variables, race membership and other grain attributes. The test revealed a highly significant association between botanical race membership and tannin (df=4, P < 0.0001) and pericarp thickness (df=4, P < 0.0001). Tannin was present in 86.4% of the accessions from the caudatum race, 83.3% of the guinea, and 71.4% of the durra-bicolor types; whereas the durra, guinea-caudatum, and kaffir races had lesser proportion (50% accessions that contain tannin. About 74% of the caudatum accessions had thick pericarp, whereas durra (33.6%) and durra-bicolor (22.8%) by thin pericarp land-races (Table 2-2).

Association between the attributes of landraces and bioclimatic variables

The result for the spearman rank correlation analysis is presented in Table 2-6. Tannin presence was positively associated with levels of annual precipitation (r=0.15, P< 0.001), and precipitation during the wettest quarter (r=0.15, P< 0.001), and precipitation during the coldest quarter (r=0.217, P< 0.0001). About 61% of the accessions contained tannin, but the distribution of tannin sorghums varies across regions and bioclimatic zones. Tannin types tend to dominate the wetter regions in the Western part of the country, while drier regions in the East and Northeast had less proportion of tannin sorghums (Figure 2-6). Likewise, landraces originating from dryer areas of the country tend to possess compact panicles. Panicle compactness generally had a negative

correlation with annual precipitation (r=-0.247, P< 0.0001), precipitation during the wettest quarter (r=-0.18, P< 0.0001), and precipitation during the coldest quarter (r=-0.286, P< 0.0001). However, compact panicle was positively associated with precipitation during the warmest quarter (r=0.185, P< 0.0001). Like panicles, large-seeded landraces were associated with dry agroclimate; that is, accessions grown in the drier environment tend to have larger seed sizes. HKW was negatively correlated with annual precipitation (r= -0.247, P< 0.0001), precipitation during the wettest quarter (r=-0.161, P< 0.0001) and precipitation during the coldest quarter (r= -0.194, P< 0.0001). However, it was positively correlated with precipitation during the warmest quarter (r= 0.185, P< 0.0001). Kernel width also had a similar pattern of association with bioclimatic variables as HKW (Table 2-5). Pericarp thickness had a different pattern of association with precipitation; it was negatively associated with precipitation during the driest month (r= -0.175, P< 0.001), and during the driest quarter (r= -0.216, P< 0.0001), and warmest quarter (-0.211, P< 0.0001). It was positively correlated with precipitation during the wettest quarter (r=0.129, P< 0.001) and coldest quarter (r=0.187, P< 0.0001). We also evaluated the correlation of these plant attributes with monthly precipitation. Precipitation during the months of September and October were significantly correlated with panicle compactness (September-r=-0.18, October= -0.19), tannin (September-r=0.18, P< 0.0001; October- r=0.19, P< 0.0001). Months September and October, mostly grain-filling periods for most sorghum genotypes, coincide with the period of active grain and panicle infection by grain molding pathogens.

PCA on combined bio-climate variables is shown in (Figure 2-7). The plot of variables on the first and second PCs shows that landraces that originated in high precipitation areas, specifically from high coldest quarter precipitation, tend to be tannin type. Landraces with compact panicles, more translucent, and larger mass arise from drier areas. Pericarp thickness in the first two dimensions tends to cluster with precipitation seasonality and temperature seasonality. Additionally, traits that were positively correlated with precipitation of origin had positive association with grain mold resistance. For example, tannin sorghums which originated in high precipitation areas were more tolerant while non-tannin, compact, large kernel genotypes tend to be susceptible to the disease when grown in disease hot-spot areas.

Caudatum race is mainly distributed in humid lowland areas of the country (50.5%). Durra is distributed across wet-highland (40.1%) and dry lowland (24.1 %) areas. Durra-bicolor is mainly

distributed in wet-highland and wet intermediate altitude regions. Bicolor is also mainly distributed across wet highlands while guinea-caudatum is distributed across most of the traditional geographic zones. The caudatum landraces mostly occur in humid warmer areas (Annual mean Temperature=22.3°C, and annual precipitation=1043 mm), while durras are concentrated in colder but drier regions (Annual mean Temperature=20.6 and Annual precipitation=848 mm). Durra-bicolor is more abundant in colder areas of the country.

Genomic support for the role of grain and panicle attributes for adaptation

Since the first few PCs of genome wide PCA were responsible to discern botanical races, they were used as a proxy for testing adaptation. To test whether these PCs were related to bioclimatic variables, correlation analysis was conducted using spearman method. The first genomic PC was correlated with precipitation variables *i.e* annual precipitation ($r=-0.283$, $P < 0.0001$), and precipitation during the coldest quarter ($r=-0.285$, $P < 0.0001$). The second genomic PC correlated with annual temperature ($r=-0.344$, $P< 0.0001$), isothermality ($r=0.289$, $P < 0.0001$), temperature seasonality (-0.248, $P < 0.0001$), temperature of the warmest month ($r= -0.366$, $P < 0.0001$), minimum temperature of the coldest month ($r = -0.336$, $P < 0.0001$), mean temperature of the wettest quarter ($r= -0.303$, $P < 0.0001$), precipitation during the driest month ($r=0.251$, $P < 0.0001$), precipitation seasonality ($r=-0.215$, $P < 0.0001$), precipitation during the driest quarter ($r=0.335$, $P < 0.0001$), precipitation in the coldest quarter ($r= -0.244$, $P < 0.0001$). The third Genome wide PC of Ethiopian collection is correlated with annual precipitation ($r=0.276$, $P < 0.0001$), precipitation of the coldest quarter ($r=0.256$, $P < 0.0001$) (Table 2-7).

Correlation between genomic PCs and grain attributes (Table 2-8) showed the first PC to be associated with several grain and panicle attributes including seed mass ($r=0.444$, $P < 0.0001$), head compactness ($r=0.413$, $P < 0.0001$), susceptibility to grain mold ($r=0.36$, $P < 0.0001$), kernel width ($r=0.464$, $P < 0.0001$), kernel thickness ($r=0.403$, $P < 0.0001$), and tannin presence ($r=-0.335$, $P < 0.0001$). PC was associated with pericarp thickness ($r=-0.359$, $P < 0.0001$), kernel length ($r=-0.256$, $P < 0.0001$). The fourth PC was associated with kernel translucence ($r=0.267$, $P < 0.0001$).

The top 0.1 % outlier SNPs which contributed the most to PCs 1, 2, 3,4, and 6 were further analyzed for their association with known genes. A total of 446, 326, 327, 361, and 416 genes were identified within the 50 kbp region flanking the outlier SNPs. As the PCs were correlated with plant attributes and adaptive traits, SNPs in the neighboring region were further scanned for the presence of known grain, panicle, and adaptation-related *priori* genes (Table 2-9, Figure 2-8). Ortholog genes attributed to variations in inflorescence, grain weight, and tannin biosynthesis were identified near the outlier SNPs. Moreover, genes related to heat shock and cold shock tolerance were also placed across the PC axis, discriminating durra-bicolor from caudatum or durra (PC2).

GWAS of plant attributes

MLM and BLINK did not produce SNPs with significant association to HKW. However, Farmcpu yielded 7 SNPs (FDR< 0.05) Table 2-10. Of these SNPs, SNP_1_73446733 (SORBI_3001G458400) was found to be associated with the beta-glucosidase gene. Similar to HKW, MLM did not yield a significant association for panicle compactness. BLINK nevertheless produced 4 SNPs with significant association (FDR < 0.05). Out of these, SNP_5_6993037 and SNP_6_55438950 had significant associations to genes SORBI_3005G063700 and SORBI_3006G203400, respectively, where the former codes for a protein similar to F-box/WD-40 repeat-protein and the later is related to growth-regulating factor 3.

MLM did produce a significant association with tannin and kernel translucence. SNP_4_6231642 was significantly associated with tannin presence and was located within the gene transparent-testa (maf=0.458, FDR < 0.0001). Kernel Translucence was also associated with SNP_6_46657114 (maf=0.121, FDR=0.012) which is near (~100 bp) glutamate decarboxylase 2 gene. Association analysis using BLINK also yielded a significant connection with SNP SNP_6_46696897 (maf= 0.199, FDR < 0.0001) which is near the gene ZEAXANTHIN EPOXIDASE (SORBI_3006G097500).

Genome Environment Association (GEA)

We conducted GEA to identify genes that are associated with adaptation across precipitation gradients. GEA using MLM did not identify significant (FDR < 0.05) SNPs.

However, SNP_1_16432372 was found to be associated (BLINK-FDR < 0.0001, FARMCPU-FDR < 0.01, maf =0.228) with precipitation. The SNP is within the gene coding for 4-*coumarate-CoA ligase* like gene.

Another precipitation variable that saw significant association is the October month precipitation. The SNP _3_60581629 was significantly associated (FDR < 0.001, maf =0.228) with the October month precipitation. The SNP is within the gene transcript of mitogen-activated protein kinase 3. Other mitogen-activated protein kinase genes were also identified at 8, 14, and 49 kb from the SNP. Precipitation for September month was also associated with SNPs linked with genes such as *senescence-associated gene 20*, *ethylene-responsive transcription factor ERF014* and *CHALCONE SYNTHASE 1* (Table 2-11).

Discussion

Sorghum is one of the five most important crops in the world. It is the second most widely used feed and food grain in the US and sub-Saharan Africa. However, sorghum has constraints that limit its value as animal feed and human food. Although it has similar and even better protein content, the digestibility of sorghum proteins based on pepsin assay is lower than most cereals. Even if the significance of tannin as one of the anti-nutritional factors is diminishing due to the exclusive use of non-tannin genotypes, the impediment to protein availability, recalcitrant protein, and starch are still constraints undermining the biological value of sorghum. Researchers have speculated the nature of sorghum grain development and maturation may have contributed to the low bioavailability of its nutrients. Sorghum is one of the few crop species that develop and mature naked and exposed to climatic and biotic agents. But little information is available on the impact of this on grain biochemical properties and the response mechanisms that the crop may have developed to cope under such environments. However, sorghum and other naked grains appear to have distinct grain properties that may be regarded as adaptation mechanisms. In this study, we hypothesized that grain structural properties (adaptive traits) that appear to protect its naked kernel may be among the factors shaping the overall sorghum grain attribute that may be a culprit for its compromised nutritional value.

The Ethiopian sorghum collection offers a unique opportunity to study the adaptive traits that shaped grain phyisco-chemical attributes. Due to large bioclimatic variations among major

sorghum-producing areas of the country, the landraces are subjected to an array of selection forces that eventually led to the development and evolution of unique sorghums adapted to their unique agroecology. Long-term selection against the biotic and abiotic factors has been shown to have led to the divergence in various morphological attributes between the landraces (Lasky et al., 2015; Wang et al., 2020a).

The distinct morphological features and geographical distribution of the botanical races had long been reported as one face of the manifestation of millennia-old sorghum adaptation (Stemler et al., 1975). The Ethiopian core collection under this study is composed of all the major botanical races as expected was dominated by durra and caudatum races (Table 2-1). Guinea and kaffir were a minority in the collection. All the major races of sorghum have been reported to be present in Ethiopia (Doggett, 1970; Harlan and de Wet, 1972; Harlan and Stemler, 1976; Menamo et al., 2021). Few other authors reported kaffir as part of Ethiopian germplasm (Subramanian et al., 1995; Tirfessa et al., 2020). From the intermediate races, durra-bicolor and guinea-caudatum appear as significant groups suggesting an ecological overlap between the two races. These two were also reported as important intermediate races in ICRISAT Ethiopian collections (Reddy et al., 2002). Durra-caudatum is another significant intermediate race in the Ethiopian-sourced ICRISAT collection. As both durra and caudatum are dominant in Ethiopia, durra-caudatum was expected to be an important intermediate race. However, Nevertheless, we did not identify any durra-caudatum in our result. One explanation may be that the Ethiopian durra-caudatum was not represented in the reference set. As the Ethiopian durra is relatively distant from durras of Indian and Sudanese origin, the durra caudatum landraces used as a reference for the supervised analysis may not genetically represent the prevalent durra-caudatum in Ethiopia.

Grain and panicle attributes are associated with the botanical races and agroclimatic cues. The major constraints of disease and pests, including insects, are associated with climatic factors, and thus a response to these biotic agents may be another driver for adaptation. Grain mold, a critical disease severely constraining grain size, germination, and overall fitness of sorghum, is caused by multiple species of fungal pathogens (Thakur et al., 2008). It is particularly dominant in high temperature and humid areas (Ackerman et al., 2021). The grain and panicle characteristics are significant attributes conditioning plant response to biotic and abiotic factors. Sorghums with compact-panicle were often found susceptible to mold and other grain/panicle diseases. Compact

architecture favors moisture accumulation and delays drying, creating an ideal condition for fungal invasion (Thakur et al., 2008; Sharma et al., 2010). This also correlates with insect damage (Sharma et al., 1994). The presence of tannin and phenolic compounds in sorghum grain had been linked to resistance against grain mold (Bandyopadhyay et al., 1988; Ackerman et al., 2021).

Correlation analysis between bioclimatic variables and sorghum grain attributes showed that tannin and pericarp thickness were positively correlated with precipitation variables, confirming the general observation that the tannin trait tends to be common in landraces from high rainfall regions. These regions are also hot spots for grain molding pathogens which implies that farmers of the regions were perhaps after grain mold resistance and bird tolerance, not the tannin trait per se. Like tannin, pericarp thickness was also positively correlated with precipitation with rainfall. The correlation pattern to monthly precipitation is like that of tannin, and this may be associated to grain mold resistance. But there are contradictory reports in the literature. As thick pericarp is due to the accumulation of starch at the pericarp's meso layer, some authors argue that it is deemed to produce a favorable environment for fungal growth (Glueck et al., 1980). However, Esele et al. (1993), using the RIL population, showed no relationship between pericarp thickness and grain mold resistance. In maize, pericarp thickness was associated with Phlobaphene concentration which is associated with resistance to fungal attack (Ackerman et al., 2021). The current result indicates that thick pericarp genotypes are favored in high precipitation areas. The correlation may not be because thick pericarp renders resistance to grain mold; it may be due to co-selection with other traits imparting resistance, especially tannin, which occurs in thick pericarp genotypes. In this study, for tannin-free accessions, we noted very weak and statistically insignificant correlation between pericarp thickness and precipitation ($r=0.075$, $P=0.2213$), while this correlation in tannin sorghums was stronger ($r=0.25$, $P < 0.0001$). Likewise, the association between precipitation and tannin became stronger (changed from $r=0.18$ for thin pericarp genotypes to $r=0.25$ for thicker pericarp genotypes). This suggests that thick pericarp and tannin presence were co-selected in disease prone high precipitation areas. All in all, the association between pericarp thickness and precipitation does not seem to be due to the thick pericarp imparting resistance to grain mold by itself but because of its correlation with tannin and other traits. The question of why tannin presence and pericarp thickness traits are correlated needs an answer. In fact, though not represented in the study accessions, there are unique variants of sorghum cultivars called *bobe* adapted to the warm and humid lowlands of Western Ethiopia.

These cultivars are unique in that they possess large seeds with thick pericarp, soft-kernels, and are resistant to foliar and panicle diseases. The key driver for the adaptation of such cultivars by the *gumuz* community of the area is its processing attributes, ease of manual grinding of the grains, and quality of porridge.

Just like tannin, panicle structure also appears to mediate crop adaptation, especially grain mold resistance. Analysis of the data from two disease prone environments, Bako and Pawe in western Ethiopia, revealed that correlation between tannin presence and grain mold incidence was highly significant both for Bako ($r=- -0.35$, $P< 0.0001$) and Pawe ($r=-0.25$, $P< 0.0001$). There was similar correlation between grain mold susceptibility and panicle compactness at both Bako ($r=0.26$, $P< 0.0001$) and Pawe ($r=0.28$, $P< 0.0001$). The role of grain size in shaping adaptation may not be direct and some hypothetical. Wang et al. (2020) presented genomic support for the adaptation of larger seeds to drier environment using the colocalization of grain weight loci with precipitation gradient. Some argue that larger seeds have good emergence and excellent seedling vigor that facilitate crop establishment to help withstand drought stress. In our study, the main cause for the large seed accessions dominating the dry areas is likely due to the drought tolerance of the durra race which is generally large seeded and has compact head. Likewise, large-seeded durra is not common in high precipitation regions due to the associated compact head that favors grain mold incidence. Research needs to untangle the large seed trait from compact panicle if large seeded open panicle types are needed in high rainfall areas.

Given the fact that grain weight parameter is associated to population structure, and hence to the PCs controlling for population structure, GWAS using MLM failed to identify SNPs associated with the trait. Wang et al. (2020) also reported similar result using a global diversity population where controlling for population structure failed to identify genomic loci associated with grain weight. However, FARMCPU identified an SNP (SNP_1_73446733) associated with beta-glucosidase. In watermelon, the beta-glucosidase gene function was shown to have a significant association between seed size and weight. Using the expression atlas from Uniprot (https://www.uniprot.org/database/DB-0004), the gene is also highly expressed in sorghum during and after flowering in the seeds. Another gene, SORBI_3003G291200 associated with SNP_3_62350910 had its Arabidopsis ortholog annotated as auxin response factor2 gene. In

Arabidopsis, the gene is responsible for integrating auxin signaling with developmental processes, including the size of seeds (Schruff et al., 2006).

The distribution of botanical races across the country appears to be influenced by bioclimatic variables. The compact durra races mainly occur in dryland and sub-humid high-land areas of the country (Stemler et al., 1975, 1977). They are not preferred in high rainfall areas because of their compact panicle and largely tannin-free or low tannin content- conducive traits for grain mold diseases. Only a tiny proportion of durra sorghum has tannin, and the majority are goose-necked compact panicles. Moreover, unlike other races, the glume covers less than half of the grain and is directly exposed to biotic and abiotic agents. Durra-bicolor races are mostly dominant across cooler and high precipitation areas, perhaps due to the relatively open panicles. However, the caudatum race was grown in somewhat higher temperature and precipitation regions. However, the prevalence of caudatum landraces in relatively higher precipitation areas contradicts with Stemler et al. (1977) where caudatum was reported to be restricted to hot dry land savannas. Most cuadatum landraces from the Ethiopian collection contain tannin, a character that perhaps imparted their adaptation to high precipitation region in the country (Table 2-2).

PCA was utilized for the detection of genomic signatures of selection (Duforet-Frebourg et al., 2016; Luu et al., 2017). In this study, PCs which discriminated the botanical races were also correlated to precipitation gradient and grain attributes, supporting that grain attributes as adaptive traits. Moreover, some of the outlier SNPs associated with the PCs had been linked to known *priori* genes related to inflorescence and grain weight. PC1 resolved durra from the rest of the botanical races, and the gene SORBI_3001G468400 is an ortholog to the maize *Prol1.1* codes for the basic helix-loop-helix transcription factor. The gene had been previously shown to be related to multiple domestication traits and controlled the number and size of inflorescence and kernel weight (Wills et al., 2013). Other *priori* genes associated with this PC are two *GS2/GL2* paralogs, SORBI_3004G269900 & SORBI_3006G203400, homologs to the rice *GS2*, which codes for rice *OsGRF4* (growth-regulating factor 4) (Li et al., 2016). PC1 was also correlated with SORBI_3003G286500 ortholog to the inflorescence *priori* gene sparse inflorescence1 (*spi1*). The gene codes for flavin-binding monooxygenase family protein pivotal in auxin biosynthesis and is involved in lateral and axillary meristem organ formation in maize (Gallavotti et al., 2008).

The second PC from the genome wide PCA was correlated with precipitation and temperature bio variables. It is not a surprise that it is associated with temperature-related bio-variables as it discerned durra-bicolor adapted to cooler environments and caudatum mainly dominant in the hotter landscape. Post-hoc, we searched the cold tolerance-related genes in proximity to the outlier SNPs. We found the uncharacterized gene SORBI_3006G228000, which possesses a cold-shock domain and may need further investigation.

GEA revealed significant associations of SNPs with sorghum adaptation across precipitation variables. Genes *4-COUMARATE-COA LIGASE LIKE 1* and *CHALCONE SYNTHASE 1* are both involved in phenolic secondary metabolite synthesis, which is involved in many aspects of plant physiology, including disease resistance. COUMARATE-COA LIGASE LIKE 1 is a ligase involved in lignin biosynthesis. CHALCONE SYNTHASE is the first committed step in flavonoid biosynthesis tasked with the conversion of 4-coumaroyl-CoA to naringenine chalcone. Both enzymes are involved in polypropanoid pathway where COUMARATE-COA LIGASE LIKE 1 is responsible for primary defense by forming a chemical barrier (Chezem et al., 2017) and is known to confer resistance to diseases (Liu et al., 2017). Apart from its importance for disease resistance, its polypropanoid polymer product, lignin, has become the focus of many investigations in the ethanol production industry due to its negative role in fermentable sugar yield (Chen and Dixon, 2007). In sorghum, a paralogous gene to 4-*COUMARATE-COA LIGASE LIKE 1* was found to control the brown midrib phenotype (*bmr2)* (Saballos et al., 2012). Chalcone synthase is also involved in disease resistance, which is upregulated in response to fungal infection (Lue et al., 1989; Cui et al., 1996). The association of these loci with precipitation during post-flowering period may indicate when sorghum is vulnerable to diseases and the importance of disease-responsive genes in such conditions.

The top 1% outlier GEA SNPs using GLM were searched against *priori* genes related to grain quality traits. Genes related to anthocyanin regulation and beta and delta kafirin genes were identified to be linked with the outlier SNPs (Table 2-9). A study that involved sequencing of kafirin genes found very limited functional divergence among diverse accessions (Laidlaw et al., 2010)

Conclusion

Ethiopia has diverse agroecology, and sorghum has been cultivated in the country for millennia. This has opened an opportunity to understand how grain and panicle attributes played a role in the overall adaptation of the crop. In this study, we showed that grain and panicle traits had played a significant role in the adaptation of sorghum across agroecology. However, there are numerous plant traits that appear to be limited to only certain agroecology such as grain size in the dry areas and disease resistance in the wet areas. Future studies need to untangle some of these beneficial traits that are restricted to certain agroecology to develop cultivars that transcend across the traditional adaptation and grow in all regions of the country. The result indicates that any future endeavor targeting grain attributes should also consider the characteristics of the target environment.

Reference

Abate, M., T. Hussien, W. Bayu, and F. Reda. 2017. Screening of Ethiopian sorghum (Sorghum bicolor) landraces for their performance under Striga hermonthica-infested conditions. Plant Breeding 136(5): 652–662.

Ackerman, A., A. Wenndt, and R. Boyles. 2021. The sorghum grain mold disease complex: Pathogens, host responses, and the bioactive metabolites at play. Front Plant Sci 12.

Alexander, D.H., and K. Lange. 2011. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics 12(1): 1–6.

Alexander, D.H., S.S. Shringarpure, J. Novembre, and K. Lange. 2015. Admixture 1.3 software manual. Los Angeles: UCLA Human Genetics Software Distribution.

Bandyopadhyay, R., L.K. Mughogho, K.E.P. Rao, and others. 1988. Sources of resistance to sorghum grain molds. Plant Dis 72(6): 504–508.

Blum, A. 2004. Sorghum physiology. Physiology and biotechnology integration for plant breeding. CRC Press. p. 136–204

Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, et al. 2007. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23(19): 2633–2635.

Bullard, R.W. 1988. Characteristics of bird-resistance in agricultural crops.

Chen, F., and R.A. Dixon. 2007. Lignin modification improves fermentable sugar yields for biofuel production. Nat Biotechnol 25(7): 759–761.

Chezem, W.R., A. Memon, F.-S. Li, J.-K. Weng, and N.K. Clay. 2017. SG2-type R2R3-MYB transcription factor MYB15 controls defense-induced lignification and basal immunity in Arabidopsis. Plant Cell 29(8): 1907–1926.

Cuevas, H.E., and L.K. Prom. 2013. Assessment of molecular diversity and population structure of the Ethiopian sorghum [Sorghum bicolor (L.) Moench] germplasm collection maintained by the USDA--ARS National Plant Germplasm System using SSR markers. Genet Resour Crop Evol 60(6): 1817–1830.

Cui, Y., J. Magill, R. Frederiksen, and C. Magill. 1996. Chalcone synthase and phenylalanine ammonia-lyase mRNA levels following exposure of sorghum seedlings to three fungal pathogens. Physiol Mol Plant Pathol 49(3): 187–199.

Doggett, H. 1970. Sorghum. Sorghum.

Duforet-Frebourg, N., K. Luu, G. Laval, E. Bazin, and M.G.B. Blum. 2016. Detecting genomic signatures of natural selection with principal component analysis: application to the 1000 genomes data. Mol Biol Evol 33(4): 1082–1093.

Duodu, K.G., J.R.N. Taylor, P.S. Belton, and B.R. Hamaker. 2003. Factors affecting sorghum protein digestibility. J Cereal Sci 38(2): 117–131.

Esele, J.P., R.A. Frederiksen, F.R. Miller, and others. 1993. The association of genes controlling caryopsis traits with grain mold resistance in sorghum. Phytopathology 83(5): 490–495.

Faye, J.M., F. Maina, Z. Hu, D. Fonceka, N. Cisse, et al. 2019. Genomic signatures of adaptation to Sahelian and Soudanian climates in sorghum landraces of Senegal. Ecol Evol 9(10): 6038–6051.

Fick, S.E., and R.J. Hijmans. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. International journal of climatology 37(12): 4302–4315.

Gallavotti, A., S. Barazesh, S. Malcomber, D. Hall, D. Jackson, et al. 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 105(39): 15196–15201.

Girma, G., H. Nida, A. Seyoum, M. Mekonen, A. Nega, et al. 2019. A large-scale genome-wide association analyses of ethiopian sorghum landrace collection reveal loci associated with important traits. Front Plant Sci 10: 691.

Girma, G., H. Nida, A. Tirfessa, D. Lule, T. Bejiga, et al. 2020. A comprehensive phenotypic and genomic characterization of Ethiopian sorghum germplasm defines core collection and reveals rich genetic potential in adaptive traits. Plant Genome: e20055.

Glueck, J.A., L.W. Rooney, and others. 1980. Chemistry and structure of grain in relation to mold resistance. Proceedings of the international workshop on sorghum diseases. Hyderabad, India, 11-15 December 1978. Grain molds. p. 119–140

Gomez, M.I., A.B. Obilana, D.F. Martin, M. Madzvamuse, and E.S. Monyo. 1997. Quality evaluation of sorghum and pearl millet. ICRISAT, Patancheru, India.

Harlan, J.R., and A. Stemler. 1976. The races of sorghum in Africa. Origins of African plant domestication. De Gruyter Mouton. p. 465–478

Harlan, J.R., and J.M.J. de Wet. 1972. A simplified classification of cultivated sorghum 1. Crop Sci 12(2): 172–176.

Haussmann, B., V. Mahalakshmi, B. Reddy, N. Seetharama, C. Hash, et al. 2002. QTL mapping of stay-green in two sorghum recombinant inbred populations. Theoretical and Applied Genetics 106(1): 133–142.

Hijmans, R.J., J. Van Etten, J. Cheng, M. Mattiuzzi, M. Sumner, et al. 2015. Package 'raster.' R package 734.

Husson, F., J. Josse, S. Le, J. Mazet, and M.F. Husson. 2016. Package 'FactoMineR.' An R package 96: 698.

IBPGR, and ICRISAT. 1993. Descriptors for Sorghum (Sorghum bicolor (L) Moench). International Board for Plant Genetic Resources, Rome.

Jordan, W., F. Miller, and D. Morris. 1979. Genetic Variation in Root and Shoot Growth of Sorghum in Hydroponics 1. Crop Sci 19(4): 468–472.

Kassambara, A., and F. Mundt. 2017. Package 'factoextra.' Extract and visualize the results of multivariate data analyses 76.

Kim, S. 2015. ppcor: an R package for a fast calculation to semi-partial correlation coefficients. Commun Stat Appl Methods 22(6): 665.

Laidlaw, H.K.C., E.S. Mace, S.B. Williams, K. Sakrewski, A.M. Mudge, et al. 2010. Allelic variation of the $β$-, $γ$-and $δ$-kafirin genes in diverse Sorghum genotypes. Theoretical and Applied Genetics 121(7): 1227–1237.

Lasky, J.R., H.D. Upadhyaya, P. Ramu, S. Deshpande, C.T. Hash, et al. 2015. Genome-environment associations in sorghum landraces predict adaptive traits. Sci Adv 1(6): e1400218.

Li, S., F. Gao, K. Xie, X. Zeng, Y. Cao, et al. 2016. The OsmiR396c-OsGRF4-OsGIF1 regulatory module determines grain size and yield in rice. Plant Biotechnol J 14(11): 2134–2146.

Lipka, A.E., F. Tian, Q. Wang, J. Peiffer, M. Li, et al. 2012. GAPIT: genome association and prediction integrated tool. Bioinformatics 28(18): 2397–2399.

Liu, H., Z. Guo, F. Gu, S. Ke, D. Sun, et al. 2017. 4-Coumarate-CoA ligase-like gene OsAAE3 negatively mediates the rice blast resistance, floret development and lignin biosynthesis. Front Plant Sci 7: 2041.

Lue, W.L., D. Kuhn, and R.L. Nicholson. 1989. Chalcone synthase activity in sorghum mesocotyls inoculated with Colletotrichum graminicola. Physiol Mol Plant Pathol 35(5): 413–422.

Luu, K., E. Bazin, and M.G.B. Blum. 2017. pcadapt: an R package to perform genome scans for selection based on principal component analysis. Mol Ecol Resour 17(1): 67–77.

Maina, F., S. Bouchet, S.R. Marla, Z. Hu, J. Wang, et al. 2018. Population genomics of sorghum (Sorghum bicolor) across diverse agroclimatic zones of Niger. Genome 61(4): 223–232.

Marla, S.R., G. Burow, R. Chopra, C. Hayes, M.O. Olatoye, et al. 2019. Genetic architecture of chilling tolerance in sorghum dissected with a nested association mapping population. G3: Genes, Genomes, Genetics 9(12): 4045–4057.

Melake-Berhan, A., L.G. Butler, G. Ejeta, and A. Menkir. 1996. Grain mold resistance and polyphenol accumulation in sorghum. J Agric Food Chem 44(8): 2428–2434.

Menamo, T., B. Kassahun, A.K. Borrell, D.R. Jordan, Y. Tao, et al. 2021. Genetic diversity of Ethiopian sorghum reveals signatures of climatic adaptation. Theoretical and Applied Genetics 134(2): 731–742.

Mengistu, G., H. Shimelis, M. Laing, D. Lule, E. Assefa, et al. 2020. Genetic diversity assessment of sorghum (Sorghum bicolor (L.) Moench) landraces using SNP markers. South African Journal of Plant and Soil 37(3): 220–226.

Morris, G.P., P. Ramu, S.P. Deshpande, C.T. Hash, T. Shah, et al. 2013. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proceedings of the National Academy of Sciences 110(2): 453–458.

Muleta, K.T., T. Felderhoff, N. Winans, R. Walstead, J.R. Charles, et al. 2021. The recent evolutionary rescue of a staple crop depended on over half a century of global germplasm exchange. bioRxiv.

Nida, H., G. Girma, M. Mekonen, S. Lee, A. Seyoum, et al. 2019. Identification of sorghum grain mold resistance loci through genome wide association mapping. J Cereal Sci 85: 295–304.

Nida, H., G. Girma, M. Mekonen, A. Tirfessa, A. Seyoum, et al. 2021. Genome-wide association analysis reveals seed protein loci as determinants of variations in grain mold resistance in sorghum. Theoretical and Applied Genetics 134(4): 1167–1184.

Olatoye, M.O., Z. Hu, F. Maina, and G.P. Morris. 2018. Genomic signatures of adaptation to a precipitation gradient in Nigerian sorghum. G3: Genes, Genomes, Genetics 8(10): 3269–3281.

Pugh, N.A., R. Rodriguez-Herrera, R.R. Klein, P.E. Klein, and W.L. Rooney. 2017. Identification of quantitative trait loci for popping traits and kernel characteristics in sorghum grain. Crop Sci 57(4): 1999–2006.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M.A.R. Ferreira, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics 81(3): 559–575.

Reddy, V.G., N.K. Rao, B.V.S. Reddy, and K.E.P. Rao. 2002. Geographic distribution of basic and intermediate races in the world collection of sorghum germplasm. International Sorghum and Millets Newsletter 43: 15–17.

Revelle, W., and M.W. Revelle. 2015. Package 'psych.' The comprehensive R archive network 337: 338.

Rhodes, D.H., L. Hoffmann, W.L. Rooney, T.J. Herald, S. Bean, et al. 2017. Genetic architecture of kernel composition in global sorghum germplasm. BMC Genomics 18(1): 15. doi: 10.1186/s12864-016-3403-x.

Sharma, H.C., V.F. Lopez, and P. Vidyasagar. 1994. Influence of panicle compactness and host plant resistance in sequential plantings on population increase of panicle-feeding insects in Sorghum bicolor (L.) Moench. Int J Pest Manag 40(2): 216–221.

Sharma, R., V.P. Rao, H.D. Upadhyaya, V.G. Reddy, and R.P. Thakur. 2010. Resistance to grain mold and downy mildew in a mini-core collection of sorghum germplasm. Plant Dis 94(4): 439–444.

Singh, R., and J.D. Axtell. 1973. High Lysine Mutant Gene (hl that Improves Protein Quality and Biological Value of Grain Sorghum 1. Crop Sci 13(5): 535–539. doi: 10.2135/cropsci1973.0011183X001300050012x.

Stemler, A.B.L., J.R. Harlan, and J.M.J. de Wet. 1975. Evolutionary history of cultivated sorghums (Sorghum bicolor [Linn.] Moench) of Ethiopia. Bulletin of the Torrey Botanical Club: 325–333.

Stemler, A.B.L., J.R. Harlan, and J.M.J. De Wet. 1977. The sorghums of Ethiopia. Econ Bot 31(4): 446–460.

Subramanian, V., N.S. Rao, R. Jambunathan, D.S. Murty, and B.V.S. Reddy. 1995. The effect of malting on the extractability of proteins and its relationship to diastatic activity in sorghum. J Cereal Sci 21(3): 283–289.

Tao, Y., E.S. Mace, S. Tai, A. Cruickshank, B.C. Campbell, et al. 2017. Whole-genome analysis of candidate genes associated with seed size and weight in Sorghum bicolor reveals signatures of artificial selection and insights into parallel domestication in cereal crops. Front Plant Sci 8: 1237.

Teshome, A., B.R. Baum, L. Fahrig, J.K. Torrance, T.J. Arnason, et al. 1997. Sorghum [Sorghum bicolor (L.) Moench] landrace variation and classification in north Shewa and south Welo, Ethiopia. Euphytica 97(3): 255–263.

Tessema, G., H. Nida, A. Seyoum, M. Mekonen, A. Nega, et al. 2019. Ethiopian sorghum landrace SNP and phenotype data. doi: doi:/10.4231/PYQV-AT79.

Thakur, R.P., V.P. Rao, and R. Sharma. 2008. Characterization of grain mold resistant sorghum germplasm accessions for physio-morphological traits. Journal of SAT Agricultural Research 6(1): 1–7.

Thornton, T., M.P. Conomos, S. Sverdlov, E.M. Blue, C.Y.K. Cheung, et al. 2014. Estimating and adjusting for ancestry admixture in statistical methods for relatedness inference, heritability estimation, and association testing. BMC proceedings. p. 1–7

Tirfessa, A., G. McLean, E. Mace, E. van Oosterom, D. Jordan, et al. 2020. Differences in temperature response of phenological development among diverse Ethiopian sorghum genotypes are linked to racial grouping and agroecological adaptation. Crop Sci 60(2): 977–990.

Wang, J., Z. Hu, H.D. Upadhyaya, and G.P. Morris. 2020. Genomic signatures of seed mass adaptation to global precipitation gradients in sorghum. Heredity (Edinb) 124(1): 108–121.

Waniska, R.D., L.F. Hugo, and L.W. Rooney. 1992. Practical methods to determine the presence of tannins in sorghum. Journal of Applied Poultry Research 1(1): 122–128.

Wills, D.M., C.J. Whipple, S. Takuno, L.E. Kursel, L.M. Shannon, et al. 2013. From many, one: genetic control of prolificacy during maize domestication. PLoS Genet 9(6): e1003604.

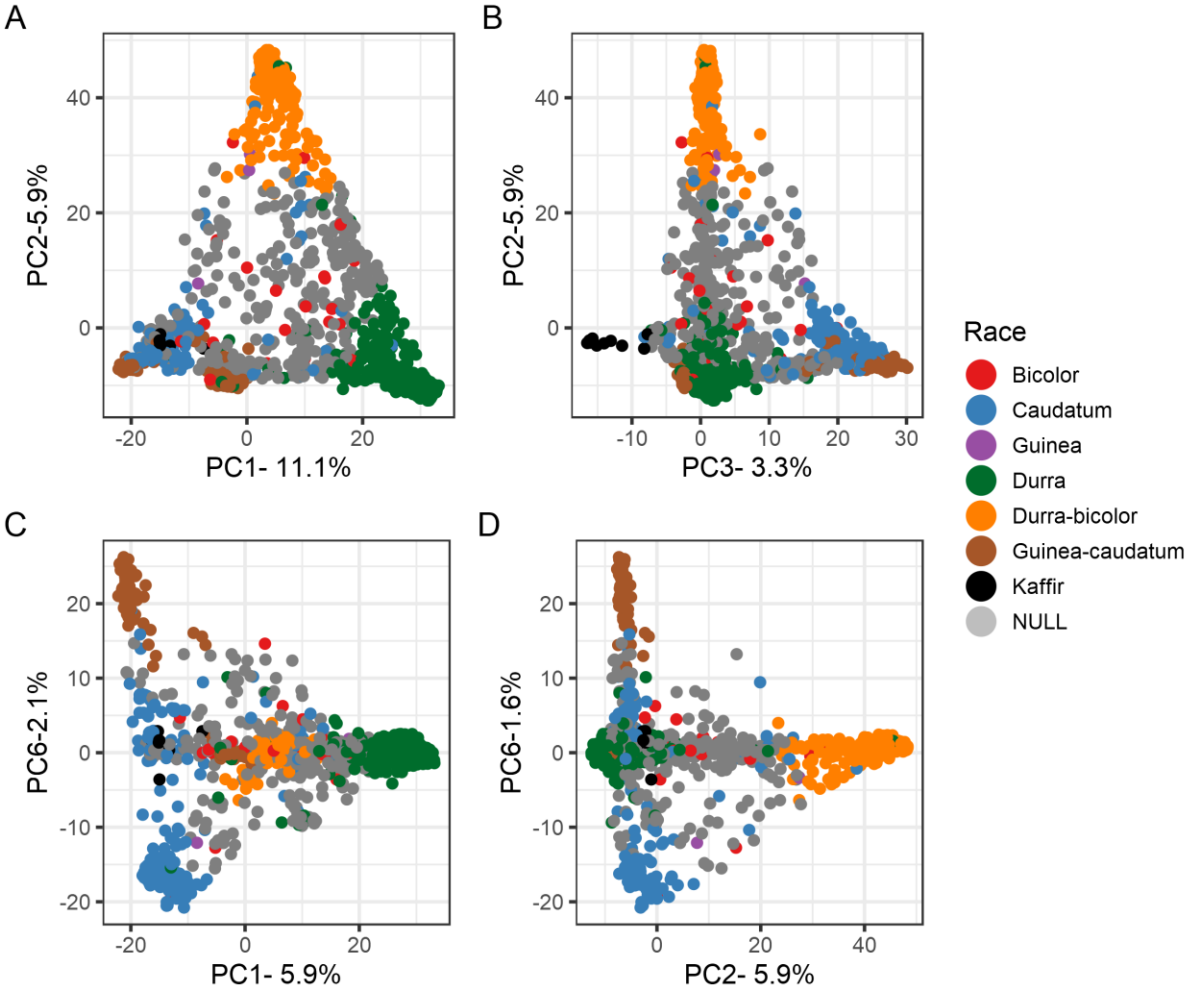


Figure 2-1 Principal component analysis of Ethiopian core collection with the genomic data where accessions are displayed against the first few PCs.

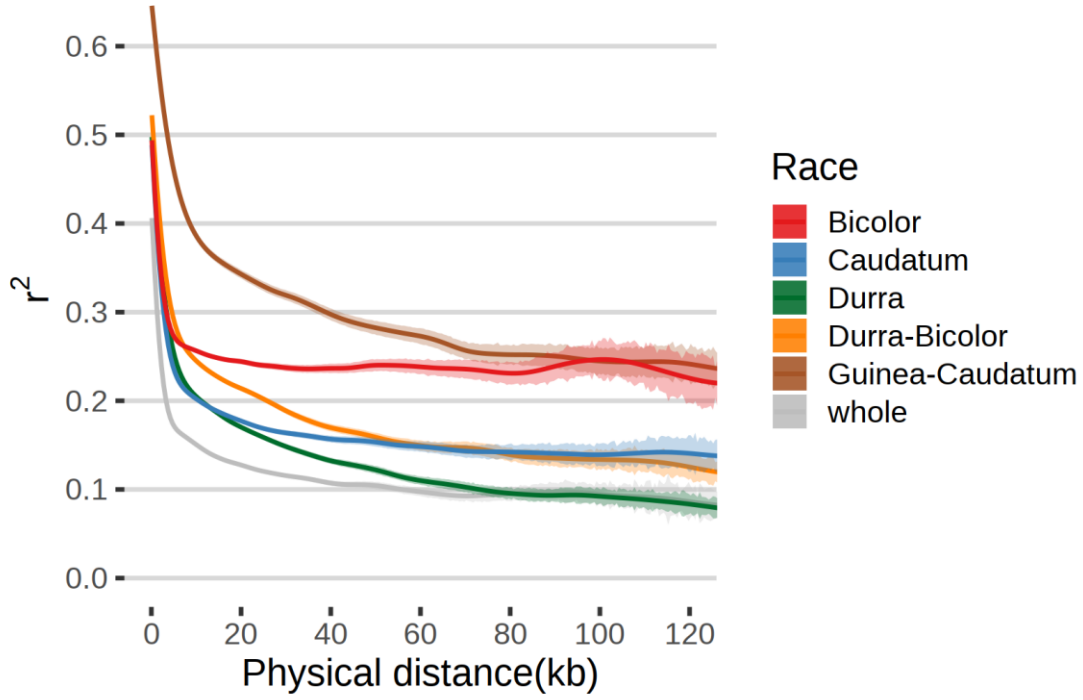


Figure 2-2. Linkage disequilibrium decay along the genome of sorghum botanical races.

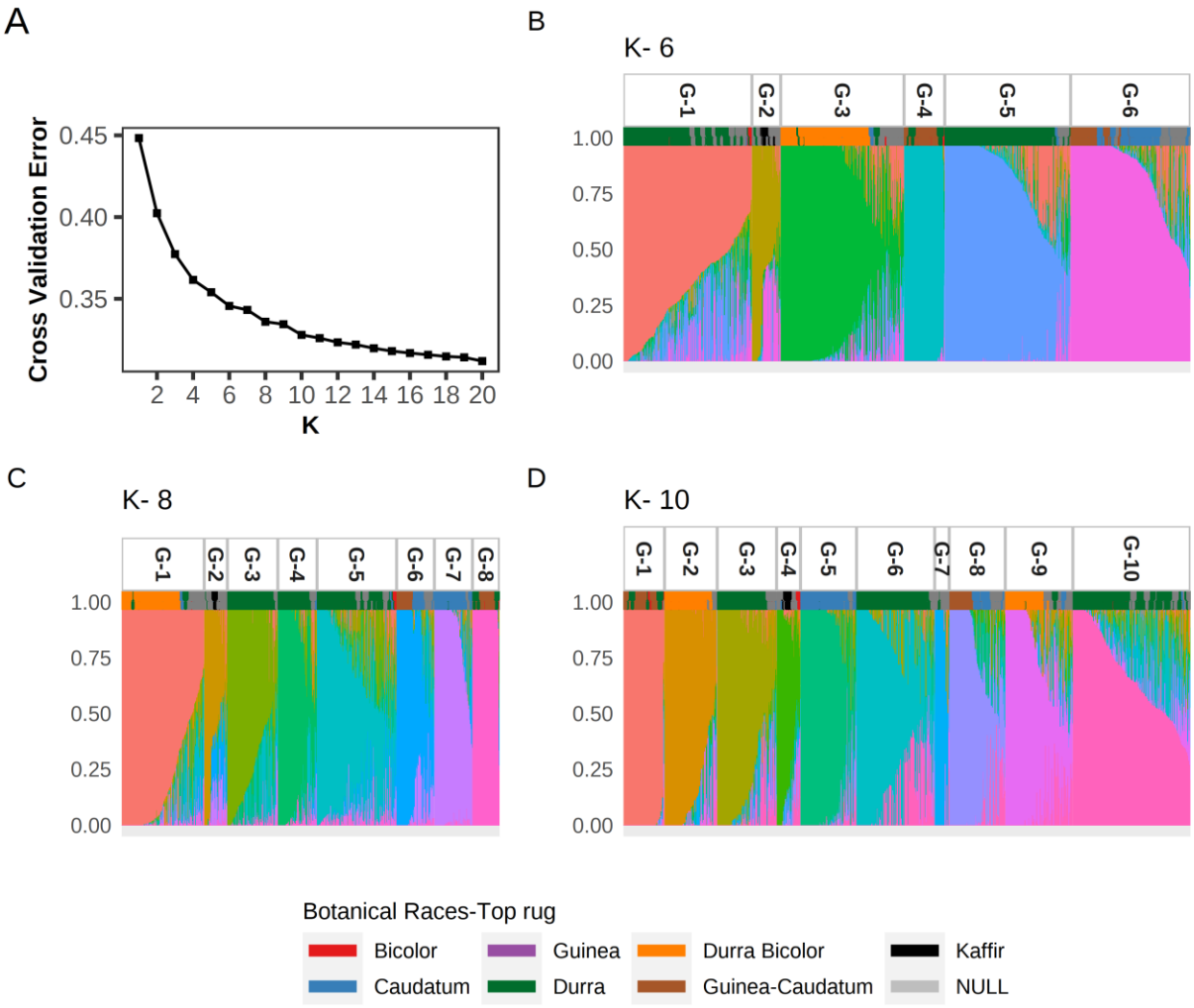


Figure 2-3 Admixture proportion of Ethiopian landraces for K=6 and K=12. (A) Cross validation error for different K values. (B-D) admixture proportions of different predicted ancestral populations. Top rug shows the assigned botanical race of the individual. The bottom bar graph shows the admixture proportions.

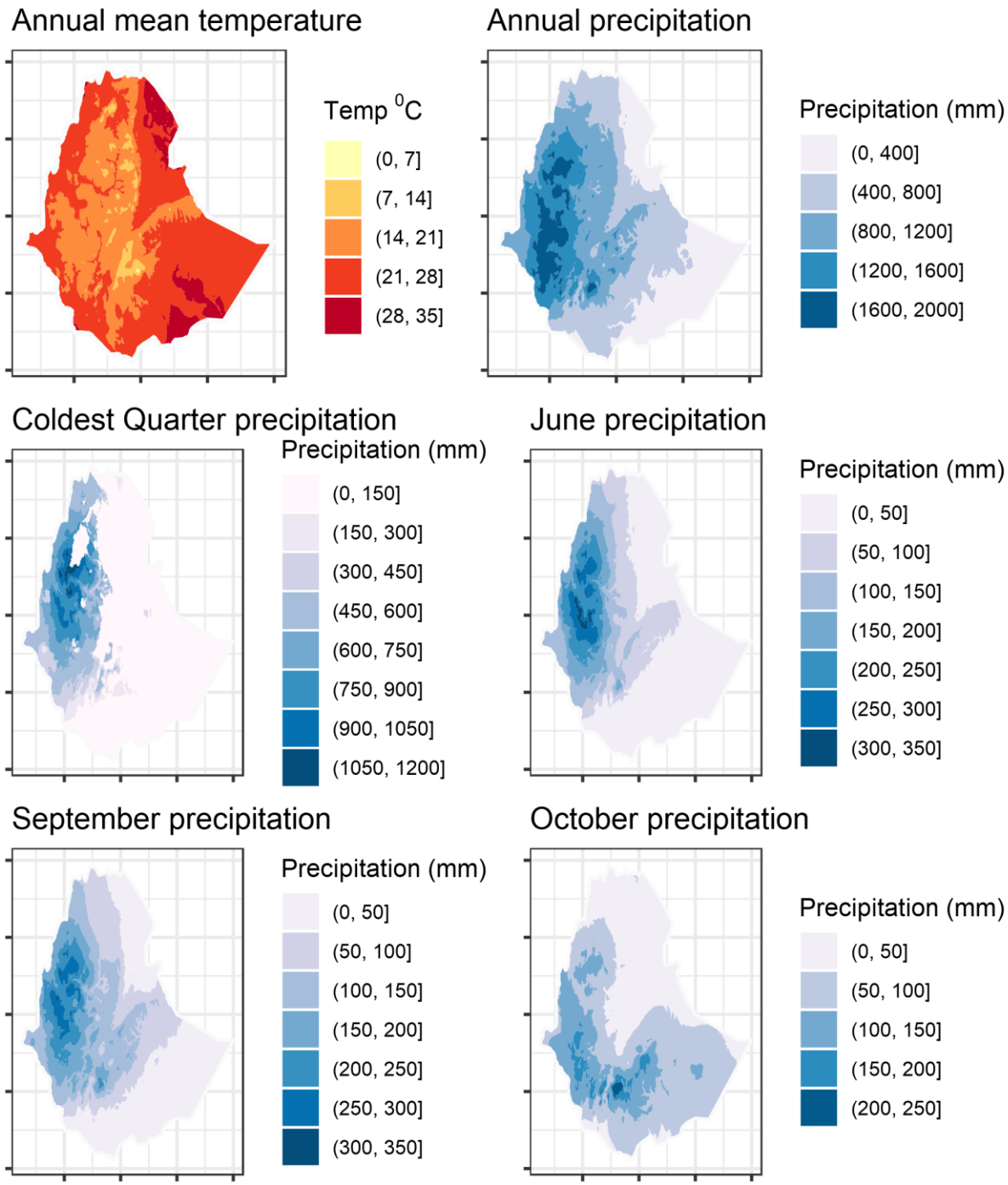


Figure 2-4 Spatial distribution of some climatic variables across Ethiopia.

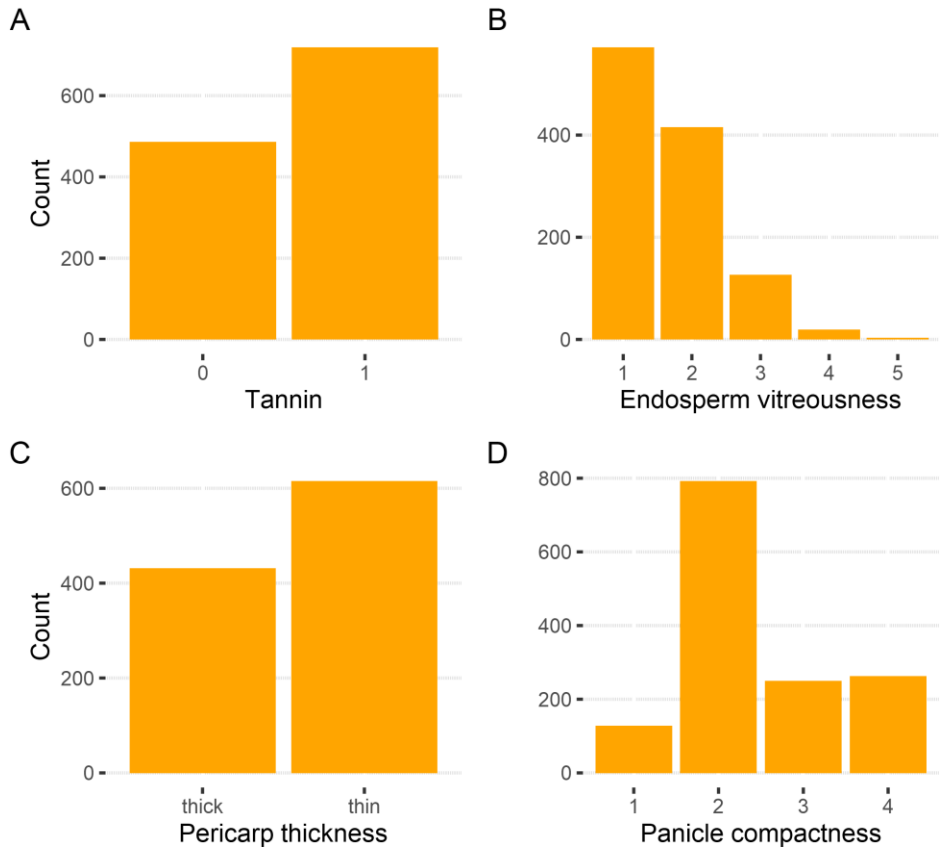


Figure 2-5. Frequency distribution of categorical grain attributes.

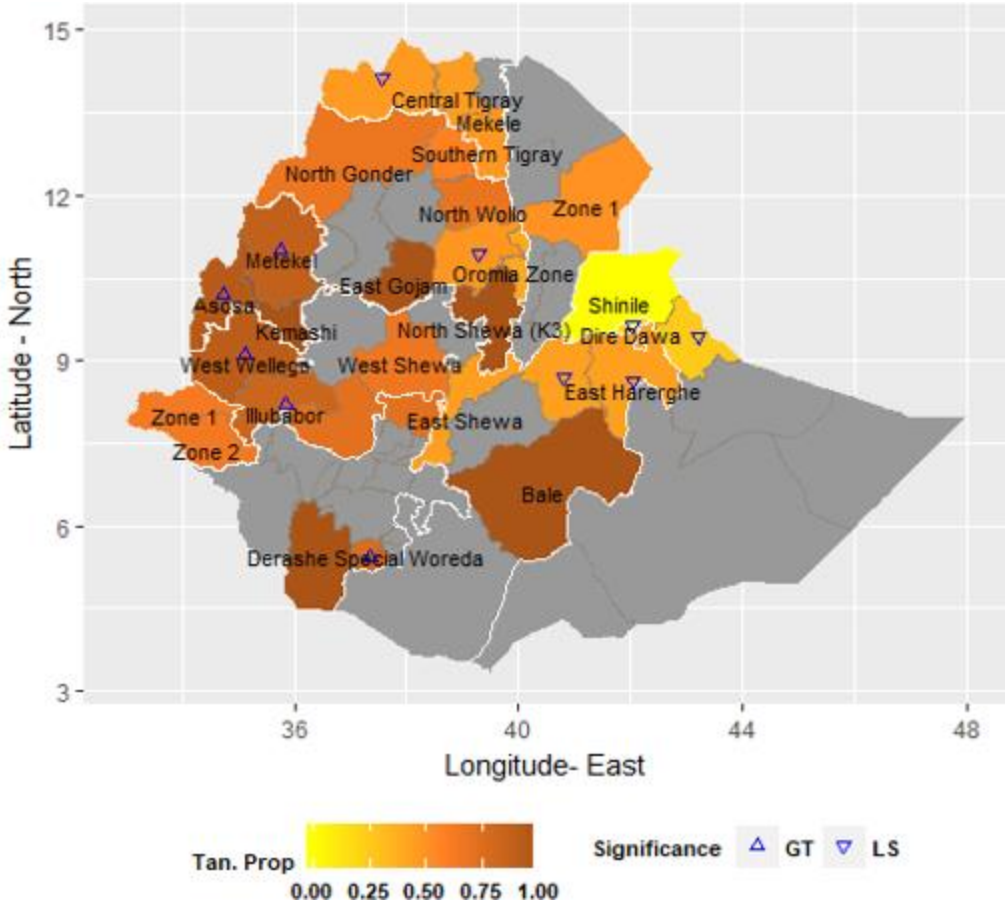


Figure 2-6 Tannin sorghum distribution across administrative zones of Ethiopia.

Triangles and inverted Triangles show significant (P< 0.05) upward and downward deviations, respectively, from the national average of tannin sorghums percentage (62%).

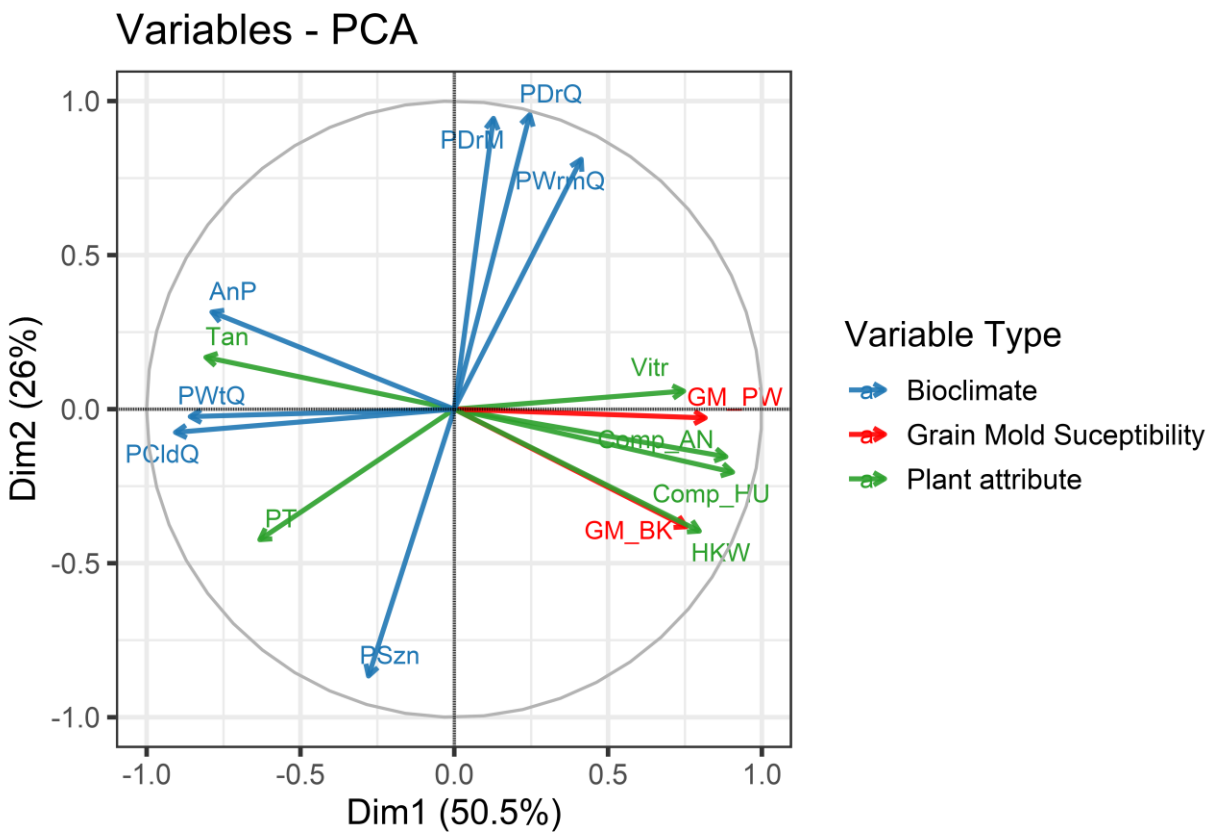


Figure 2-7 Direction of relationships for plant and bioclimatic variables across the first two dimensions (Principal components).

An-Annual, M-Month, Q-Quarter, T-Temperature, P-Precipitation, Wrm-Warmest, C-Panicle compactness, Cld-Coldest, Wt-Wettest, Dr-Driest, Min-Minimum, Mn-Mean, S-Seasonality, Dim Dimension axis, HKW-Hundred kernel weight, C- Panicle compactness, Vitr-endosperm vitreousness, GM- Grain mold sucesceptiblity, PT-Pericarp thickness, HU-Haramaya University (site), AN-Arisi Negele (site), BK-Bako (site), PW-Pawe (site).

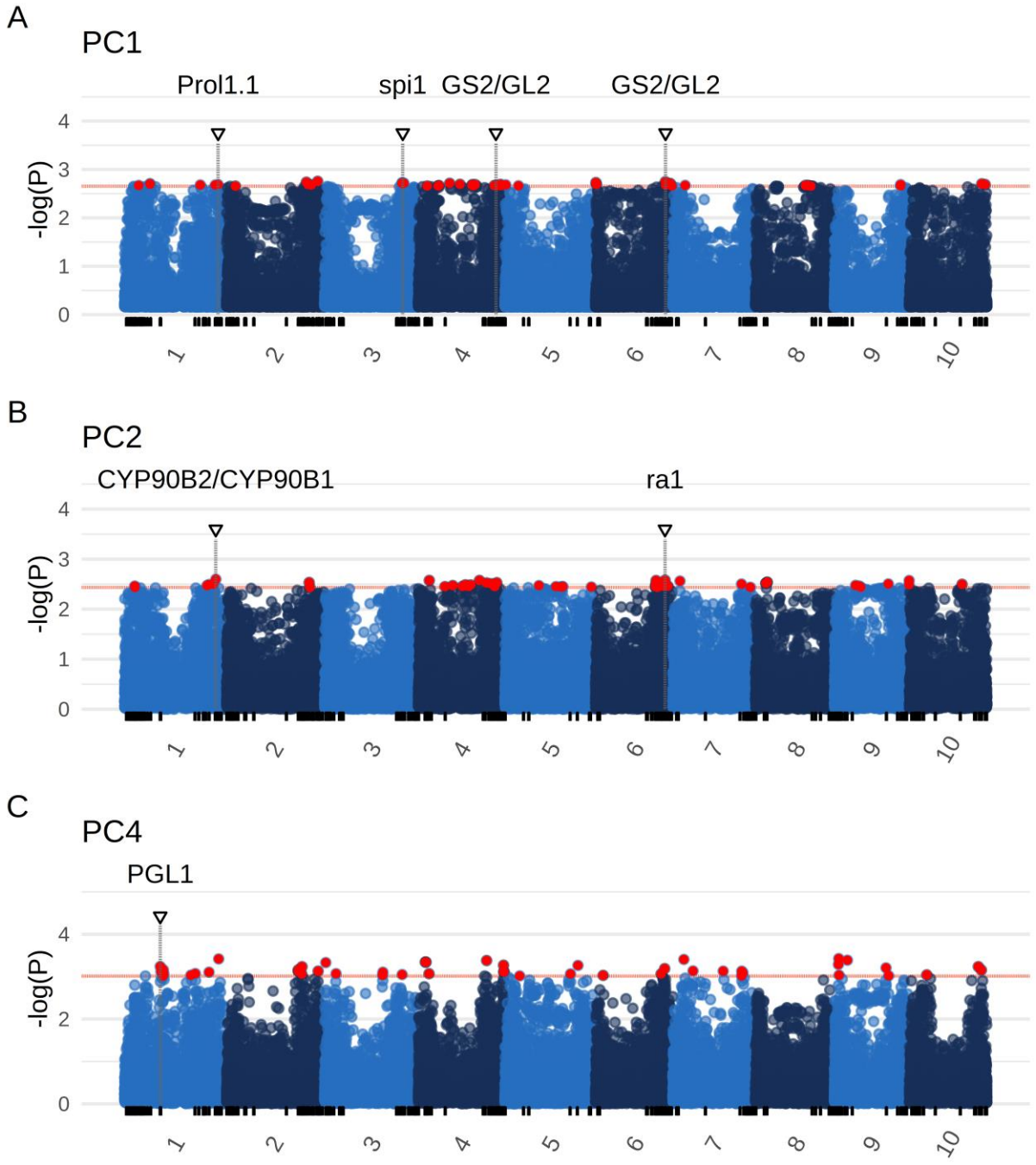


Figure 2-8 Genome scan of the loadings of the first few principal components.

The x-axis represent the position of each SNP on the chromosomes and the y-axis the -log10(P-values) using adjusted chi-square distribution. The horizontal line represents 99.9% percentile threshold. The bottom rug shows grain and panicle *priori* candidate genes, and the vertical lines show priori genes linked (50 kbp) with outlier SNPs.

Table 2-1. Race assignment of Ethiopian landraces.

Botanical Races	Reference[1]			Inferred	Race Assigned Ethiopian Core
	Ethiopian Core Collection[2]	Global Collection	Reference Total	Ethiopian Core	
Bicolor	35	84	119	0	35
Caudatum	47	362	409	165	212
Caudatum bicolor	0	77	77	0	0
Durra	80	258	338	605	685
Durra-bicolor	0	126	126	248	248
Durra vaudatum	0	157	157	0	0
Guinea	6	445	451	0	6
Guinea-caudatum	0	192	192	183	183
Kaffir	7	38	45	3	10
Total	175	1739	1914	1204	1379

[1]Reference accessions used for supervised race membership analysis using ADMIXTURE software. [2]Manually assigned.

Table 2-2 Mean performance of various plant attributes aggregated by botanical races and geographic zones.

Category	No	Accessions with tannin	Panicle Compactness (1 -4 scale)	Thick pericarp (%)	Translucence	Vitreousness	HKW (g)	YPP (g)	L (mm)	W (mm)	H (mm)	Protein (%)
Botanical races												
Bicolor	35	71.4%	1.3	50.0%	1.5	1.7	2.2	44.1	3.8	3.4	2.5	11.1
Caudatum	192	86.2%	2.1	73.8%	1.1	1.4	2.2	33.0	3.9	3.7	2.7	10.5
Durra	617	38.2%	3.3	33.6%	1.7	1.7	2.9	57.2	4.0	4.1	2.9	11.0
Durra-bicolor	236	77.2%	1.7	22.8%	1.5	1.6	1.9	51.5	3.6	3.4	2.6	9.9
Guinea	6	83.3%	1.3	20.0%	1.4	1.8	2.2	25.5	3.8	3.6	2.5	10.3
Guinea-Caudatum	145	50.9%	2.5	39.6%	1.6	1.7	2.5	47.8	4.0	3.8	2.7	10.9
Kaffir	5	50.0%	2.4	66.7%	1.0	1.7	2.3	56.5	3.7	3.5	2.6	11.7
NA	299	77.7%	2.2	40.9%	1.4	1.6	2.4	46.9	4.0	3.7	2.7	10.4
Grand Total	1536	61.1%	2.5	37.8%	1.5	1.6	2.5	50.0	3.9	3.8	2.8	10.7
Geographic zones												
Dry highland	114	63.8%	2.4	47.4%	1.4	1.6	2.6	51.9	3.9	3.8	2.8	10.6
Dry intermediate	117	43.4%	3.1	26.5%	1.8	1.8	2.7	53.9	3.8	4.0	2.8	10.7
Dry lowland	272	55.0%	2.8	34.5%	1.7	1.7	2.7	48.7	4.0	4.0	2.8	10.9
Wet highland	530	64.3%	2.5	35.6%	1.4	1.6	2.5	54.5	3.9	3.8	2.8	10.5
Wet intermediate	163	54.5%	2.4	35.6%	1.6	1.7	2.2	49.0	3.8	3.7	2.7	10.3
Wet lowland	251	74.5%	2.2	54.7%	1.2	1.5	2.4	41.2	4.0	3.8	2.7	10.9
Total/mean	1536	61.1%	2.5	37.8%	1.5	1.6	2.5	50.0	3.9	3.8	2.8	10.7

HKW - Hundred kernel weight; YPP - Yield per panicle; L- Kernel length; H- Kernel thickness, W-Kernel width

Table 2-3 Population differentiation (Fst) estimate between pairs of botanical races drawn from Ethiopian collection.

Race 1	Race 2	Fst
Bicolor	Kafir	0.11
	Guinea-caudatum	0.25
	Guinea	0.07
	Durra-bicolor	0.22
Caudatum	Kafir	0.09
	Bicolor	0.05
	Durra-bicolor	0.28
	Guinea	0.01
	Kafir	0.1
	Guinea-caudatum	0.16
Durra	Durra-bicolor	0.28
	Guinea-caudatum	0.26
	Caudatum	0.22
	Bicolor	0.14
	Durra-bicolor	0.28
	Guinea	0.05
	Kaffir	0.16
Durra-bicolor	Kaffir	0.25
	Guinea	0.14
	Guinea-caudatum	0.32
	Kafir	0.25

Table 2-4 Summary of grain and phenological attributes of landraces evaluated at Arsi-Negele in the 2016 main growing season.

Plant feature	Number	Range		Mean
		Minimum	Maximum	
YPP (g)	1327	0	222.2	49.6
HKW (g)	1420	0.6	4.7	2.5
Grain protein Content (%)	781	7.07	14.04	10.7
Kernel Length (mm)	1395	2.6	5.1	3.9
Kernel Width (mm)	1395	2.3	5.6	2.8
Kernel Thickness (mm)	1395	1.7	4.73	2.5

YPP-Yield per panicle, HKW-Hundred kernel weight

Table 2-5 Correlation of grain and panicle attributes evaluated during the main season of 2016 at Arsi Negele, Ethiopia.

Character	parameter	Tannin	Comp	PT	Translucence	Vitreousness	HKW	YPP	L	H	W	Protein
Tannin	r	1	-0.30	0.15	-0.34	-0.29	-0.28	0.00	-0.09	-0.25	-0.29	-0.27
	- $log\ p$	-	26.72	4.38	27.32	19.87	22.80	0.00	1.87	17.97	25.00	11.45
Comp	r		1	-0.21	0.31	0.13	0.44	0.40	0.10	0.37	0.50	0.17
	- $log\ p$		-	9.65	22.57	3.26	60.81	50.29	2.27	42.55	80.72	4.10
PT	r			1	-0.36	-0.19	0.01	-0.21	0.11	0.00	-0.03	0.11
	- $log\ p$			-	30.07	8.08	0.00	8.88	1.94	0.00	0.00	0.92
Translucence	r				1	0.57	0.19	0.10	0.06	0.11	0.20	0.08
	- $log\ p$				-	98.16	8.22	1.79	0.27	2.27	8.62	0.27
Vitreousness	r					1	0.13	-0.02	0.08	0.07	0.09	0.09
	- $log\ p$					-	3.14	0.00	1.13	0.65	1.44	0.42
HKW	r						1	0.34	0.56	0.64	0.72	0.37
	- $log\ p$						-	33.63	105.24	147.20	210.10	23.53
YPP	r							1	0.13	0.24	0.28	0.01
	- $log\ p$							-	3.88	15.64	20.30	0.00
L	r								1	0.39	0.52	0.22
	- $log\ p$								-	50.75	93.71	7.20
H	r									1	0.68	0.28
	- $log\ p$									-	186.73	12.73
W	r										1	0.33
	- $log\ p$										-	18.05

HKW - Hundred kernel weight; YPP - Yield per panicle; L- Kernel length; H- Kernel thickness, W-Kernel width, Comp -panicle compactness, PT – pericarp thickness

Table 2-6 Relationship between plant attributes and precipitation related bioclimatic variables of the landraces source environments.

Plant Character	Parameter	AnP	PWtM	PDrM	PS	PWtQ	PDrQ	PWrmQ	PCldQ
Tannin	r	0.154	0.124	0.044	-0.016	0.156	0.003	-0.102	0.217
	-log (P)	3.3	1.4	0.0	0.0	3.4	0.0	0.2	8.5
Translucence	r	-0.116	-0.143	-0.008	-0.101	-0.153	0.073	0.152	-0.243
	-log (P)	0.7	2.2	0.0	0.0	2.9	0.0	2.8	10.4
Vitreousness	r	-0.101	-0.110	-0.016	-0.057	-0.122	0.027	0.097	-0.168
	-log (P)	0.0	0.4	0.0	0.0	1.0	0.0	0.0	3.8
Pericarp Thickness	r	0.066	0.108	-0.175	0.173	0.129	-0.216	-0.211	0.187
	-log (P)	0.0	0.1	3.8	3.7	1.1	7.0	6.6	4.7
Width	r	-0.225	-0.126	-0.091	0.082	-0.165	-0.088	-0.003	-0.224
	-log (P)	10.9	2.0	0.0	0.0	4.9	0.0	0.0	10.9
Length	r	-0.109	-0.055	-0.070	0.073	-0.069	-0.108	-0.108	-0.008
	-log (P)	0.9	0.0	0.0	0.0	0.0	0.9	0.9	0.0
Thickness	r	-0.182	-0.120	-0.118	0.069	-0.144	-0.096	-0.004	-0.169
	-log (P)	6.4	1.6	1.4	0.0	3.2	0.2	0.0	5.2
HKW	r	-0.247	-0.128	-0.048	0.106	-0.180	-0.071	-0.018	-0.196
	-log (P)	14.0	2.2	0.0	0.8	6.4	0.0	0.0	8.0
Compactness	r	-0.145	-0.124	-0.018	-0.050	-0.161	0.043	0.185	-0.286
	-log (P)	3.4	2.0	0.0	0.0	4.7	0.0	6.9	19.7
YPP	r	-0.032	-0.049	0.144	-0.131	-0.082	0.194	0.241	-0.194
	-log (P)	0.0	0.0	3.0	2.1	0.0	7.1	12.1	7.1
Protein	r	-0.155	-0.107	-0.105	0.049	-0.110	-0.132	-0.157	-0.097
	-log (P)	1.3	0.0	0.0	0.0	0.0	0.3	1.4	0.0

r- correlation coefficient, HKW-Hundred kernel weight, YPP- Yield per panicle, An-Annual, M-Month, Q-Quarter, P-Precipitation, Wrm-Warmest, Cld-Coldest, Wt-Wettest, Dr-Driest, Min-Minimum, Mn-Mean, S-Seasonality

Table 2-7 Correlation between bioclimatic variables and genomic PCs of Ethiopian core collection.

		PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
AnT	r	-0.003	-0.344	0.087	0.016	0.006	0.033	-0.071	-0.039	0.054	-0.014
	p	0.924	0.000	0.002	0.572	0.836	0.245	0.012	0.170	0.060	0.629
Isothrm	r	-0.092	0.289	-0.079	-0.046	0.198	0.110	0.144	-0.089	-0.065	-0.061
	p	0.001	0.000	0.005	0.107	0.000	0.000	0.000	0.002	0.024	0.033
TSzn	r	0.211	-0.248	-0.035	-0.060	-0.080	-0.059	-0.058	0.023	0.089	0.136
	p	0.000	0.000	0.226	0.034	0.005	0.038	0.041	0.427	0.002	0.000
TWrmM	r	0.014	-0.366	0.091	-0.016	-0.038	0.048	-0.104	-0.009	0.064	0.024
	p	0.624	0.000	0.001	0.569	0.189	0.091	0.000	0.753	0.024	0.406
MinTCldM	r	-0.040	-0.336	0.103	0.015	0.025	0.032	-0.056	-0.033	0.043	-0.033
	p	0.157	0.000	0.000	0.602	0.376	0.266	0.050	0.253	0.130	0.245
MnTWtQ	r	0.081	-0.303	0.035	0.047	0.023	0.067	-0.040	-0.016	0.035	0.014
	p	0.004	0.000	0.215	0.103	0.412	0.019	0.165	0.586	0.223	0.612
MnTDrQ	r	-0.095	-0.350	0.125	-0.008	0.004	0.051	-0.075	-0.036	0.043	-0.046
	p	0.001	0.000	0.000	0.772	0.888	0.074	0.009	0.209	0.132	0.109
MnTWrmQ	r	0.030	-0.354	0.077	0.006	-0.014	0.026	-0.087	-0.027	0.066	0.018
	p	0.288	0.000	0.007	0.833	0.613	0.370	0.002	0.348	0.020	0.518
MnTCldQ	r	-0.054	-0.347	0.103	0.007	0.008	0.046	-0.081	-0.051	0.045	-0.034
	p	0.058	0.000	0.000	0.813	0.779	0.104	0.005	0.071	0.117	0.231
AnP	r	-0.283	-0.050	0.276	0.137	-0.243	-0.060	-0.174	0.113	-0.060	-0.124
	p	0.000	0.082	0.000	0.000	0.000	0.034	0.000	0.000	0.036	0.000
PWtM	r	-0.159	-0.095	0.181	0.005	-0.250	-0.052	-0.158	0.103	-0.025	-0.012
	p	0.000	0.001	0.000	0.851	0.000	0.071	0.000	0.000	0.378	0.676
PDrM	r	0.013	0.251	-0.083	-0.038	0.041	0.149	0.130	0.145	-0.059	-0.012

		PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
	p	0.659	0.000	0.004	0.188	0.152	0.000	0.000	0.000	0.038	0.681
PSzn	r	0.084	-0.215	-0.003	-0.174	-0.119	-0.033	-0.135	-0.035	0.083	0.169
	p	0.003	0.000	0.930	0.000	0.000	0.246	0.000	0.223	0.003	0.000
PWtQ	r	-0.227	-0.121	0.245	0.041	-0.279	-0.069	-0.182	0.113	-0.016	-0.065
	p	0.000	0.000	0.000	0.154	0.000	0.015	0.000	0.000	0.576	0.022
PDrQ	r	0.020	0.322	-0.091	0.059	0.051	0.089	0.164	0.138	-0.061	-0.053
	p	0.485	0.000	0.001	0.038	0.076	0.002	0.000	0.000	0.032	0.063
PWrmQ	r	0.112	0.335	-0.095	0.134	-0.046	-0.034	0.069	0.075	-0.077	0.012
	p	0.000	0.000	0.001	0.000	0.109	0.232	0.016	0.008	0.007	0.675
PCldQ	r	-0.285	-0.244	0.256	-0.015	-0.194	-0.024	-0.151	0.042	-0.031	-0.065
	p	0.000	0.000	0.000	0.588	0.000	0.402	0.000	0.143	0.280	0.023

An-Annual, M-Month, Q-Quarter, T-Temperature, P-Precipitation, Wrm-Warmest, Cld-Coldest, Wt-Wettest, Dr-Driest, Min-Minimum, Mn-Mean, Szn-Seasonality, Isothrm- Isothermally, PC-Principal Component

Table 2-8. Association of grain and panicle attributes with the first ten PCs extracted from Genome-wide principal component analysis of Ethiopian landraces.

Variable		PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
HKW (BLUP)	r	0.444	-0.173	-0.091	0.101	0.080	0.108	-0.023	0.020	-0.124	0.056
	p	0.000	0.000	0.000	0.000	0.002	0.000	0.380	0.443	0.000	0.033
Compact (AN)	r	0.413	0.039	-0.061	0.277	0.146	-0.014	-0.122	-0.017	-0.006	0.069
	p	0.000	0.152	0.023	0.000	0.000	0.602	0.000	0.527	0.838	0.011
Compact (HU)	r	0.497	0.016	-0.045	0.329	0.120	0.006	-0.111	-0.005	-0.064	0.094
	p	0.000	0.532	0.087	0.000	0.000	0.822	0.000	0.858	0.014	0.000
GM (BK)	r	0.370	-0.077	-0.091	-0.035	0.006	0.149	-0.119	-0.021	0.004	0.014
	p	0.000	0.003	0.000	0.180	0.803	0.000	0.000	0.417	0.880	0.577
GM (PW)	r	0.360	0.151	-0.035	0.054	0.091	0.293	0.011	0.062	-0.041	0.044
	p	0.000	0.000	0.181	0.039	0.000	0.000	0.677	0.017	0.111	0.091
Kernel thickness	r	0.403	-0.151	-0.072	0.073	0.041	-0.015	-0.057	0.071	-0.038	-0.002
	p	0.000	0.000	0.008	0.007	0.133	0.573	0.037	0.010	0.160	0.936
Kernel Length	r	0.141	-0.265	-0.055	0.017	0.021	0.122	-0.007	-0.002	-0.076	0.006
	p	0.000	0.000	0.044	0.540	0.443	0.000	0.795	0.939	0.005	0.840
Kernel Width	r	0.464	-0.189	-0.095	0.168	0.034	0.025	-0.043	0.048	-0.127	0.053
	p	0.000	0.000	0.001	0.000	0.213	0.355	0.118	0.080	0.000	0.051
Pericarp thickness	r	-0.167	-0.359	-0.002	-0.091	-0.174	-0.159	-0.029	0.047	0.002	-0.032
	p	0.000	0.000	0.946	0.004	0.000	0.000	0.363	0.140	0.944	0.310
Tannin	r	-0.335	0.000	0.016	-0.037	-0.029	-0.138	0.081	0.095	-0.115	0.083
	p	0.000	0.994	0.600	0.206	0.322	0.000	0.006	0.001	0.000	0.005
Translucence	r	0.144	0.119	-0.023	0.267	0.147	0.033	-0.043	-0.134	0.097	-0.036
	p	0.000	0.000	0.448	0.000	0.000	0.276	0.160	0.000	0.001	0.234

Variable		PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Vitreousness	r	0.048	0.045	-0.011	0.168	0.124	0.062	0.009	-0.116	0.172	-0.077
	p	0.116	0.133	0.717	0.000	0.000	0.040	0.770	0.000	0.000	0.011

Characters in the parentheses represent test locations: AN-Arsi Negele, HU- Haramaya University, BK- Bako, PW- Pawe, H-Kernel thickness, GM- Grain Mold

Table 2-9 Priori panicle-related and grain-related genes nearest (<50kbp) to outlier SNPs with top 0.1% loadings on the respective PCs.

Trait	PC	Gene name	Gene Symbol	Molecular Function	SNP_ID	Percentile	Distance (bp)
Inflorescence							
	PC1	SORBI_3003G286500	sparse inflorescence1 (spi1)	Flavin-binding monooxygenase family protein	SNP_3_61991478	99.989%	30723
	PC2	SORBI_3006G197200	Ramosa1 (ra1)	zinc-finger protein 10	SNP_6_54995748	99.997%	21204
	PC4	SORBI_3002G184600	Ramosa3 (ra3)	Halo acid dehalogenase-like hydrolase (HAD) superfamily protein	SNP_2_56856481	99.979%	210
	PC4	SORBI_3006G201600	Aberrant Panicle Organization (APO2)	floral meristem identity control protein LEAFY (LFY)	SNP_6_55309725	99.934%	16405
Grain weight							
	PC1	SORBI_3001G468400	Prol1.1	homeobox protein 21	SNP_1_74124398	99.956%	11080
	PC1	SORBI_3004G269900	GS2/GL2	growth-regulating factor 5	SNP_4_61449408	99.931%	27581
	PC1	SORBI_3006G203400	GS2/GL2	growth-regulating factor 5	SNP_6_55427490	99.943%	16175
	PC2	SORBI_3001G445900	CYP90B2/CYP90B1	Cytochrome P450 superfamily protein	SNP_1_72308374	99.999%	41980
	PC4	SORBI_3001G254100	PGL1	basic helix-loop-helix (bHLH) DNA-binding family protein	SNP_1_28194181	99.991%	20844
Heat*							
	PC1	SORBI_3002G243200		HEAT SHOCK PROTEIN 81.4	SNP_2_63263027	99.995%	36910
	PC1	SORBI_3002G243500		HEAT SHOCK PROTEIN 81.4	SNP_2_63263027	99.995%	7080

Trait	PC	Gene name	Gene Symbol	Molecular Function	SNP_ID	Percentile	Distance (bp)
	PC1	SORBI_3003G286700		heat shock transcription factor C1	SNP_3_61991478	99.989%	355
	PC1	SORBI_3010G230600		heat shock protein 70 (Hsp 70) family protein	SNP_10_57265445	99.975%	41641
	PC2	SORBI_3002G271100		heat shock transcription factor B2A	SNP_2_65454031	99.983%	12375
	PC6	SORBI_3006G005600		heat shock protein 90.1	SNP_6_849922	99.967%	18299
Cold *							
	PC2	SORBI_3006G228000*		cold shock domain protein 1	SNP_6_57262005	99.945%	80

*Post-hoc search

Table 2-10 Genes linked to SNPs associated to grain and panicle attributes identified through GWAS.

Ensemble ID	SNPID	maf	P	FDR	Trait	Method	Annotated function	SNP distance from gene
SORBI_3005G063700	SNP_5_6993037	0.223	9.94E-08	1.01E-02	Head Compactness	BLINK	F-box/WD-40 repeat-containing protein At3g52030	20,406
SORBI_3006G203400	SNP_6_55438950	0.216	3.28E-07	1.99E-02	Head Compactness	BLINK	growth-regulating factor 3	31,449
SORBI_3003G291200	SNP_3_62350910	0.092	1.21E-08	1.83E-03	HKW	FARMCPU	auxin-responsive protein IAA6-like	41,826
SORBI_3003G290700	SNP_3_62350910	0.092	1.21E-08	1.83E-03	HKW	FARMCPU	anthocyanidin 5,3-O-glucosyltransferase	31,221
SORBI_3003G290900	SNP_3_62350910	0.092	1.21E-08	1.839E-03	HKW	FARMCPU	probable glutamate carboxypeptidase LAMP1	9,887
SORBI_3001G458400	SNP_1_73446733	0.314	1.25E-06	3.798E-02	HKW	FARMCPU	beta-glucosidase 6	19,497
SORBI_3001G458600	SNP_1_73446733	0.314	1.25E-06	3.798E-02	HKW	FARMCPU	beta-glucosidase 6	28,872
SORBI_3010G019700	SNP_10_1555964	0.390	2.15E-09	3.764E-04	Pericarp thickness	FARMCPU	cytochrome P450 704C1	45,066
SORBI_3003G010100	SNP_3_924565	0.127	9.71E-11	1.479E-05	Tannin	FARMCPU	cytochrome P450 71A1	47,575
SORBI_3003G010200	SNP_3_924565	0.127	9.71E-11	1.479E-05	Tannin	FARMCPU	cytochrome P450 71A1	44,821
SORBI_3003G010300	SNP_3_924565	0.127	9.71E-11	1.479E-05	Tannin	FARMCPU	cytochrome P450 71A1	40,094

Ensemble ID	SNPID	maf	P	FDR	Trait	Method	Annotated function	SNP distance from gene
SORBI_3003G010400	SNP_3_924565	0.127	9.71E-11	1.479E-05	Tannin	FARMCPU	ras-related protein Rab7	38,008
SORBI_3004G280800	SNP_4_62316425	0.458	1.27E-11	3.885E-06	Tannin	MLM	protein TRANSPARENT TESTA GLABRA 1	Within_gene
SORBI_3004G280800	SNP_4_62316425	0.458	3.27E-11	9.972E-06	Tannin	FARMCPU	protein TRANSPARENT TESTA GLABRA 1	Within_gene
SORBI_3004G280800	SNP_4_62316425	0.458	4.49E-09	1.369E-03	Tannin	BLINK	protein TRANSPARENT TESTA GLABRA 1	Within_gene
SORBI_3004G280800	SNP_4_62334227	0.389	3.2E-08	3.255E-03	Tannin	MLM	protein TRANSPARENT TESTA GLABRA 1	18831
SORBI_3006G097500	SNP_6_46696897	0.199	3.95E-15	1.205E-09	Translucence	BLINK	zeaxanthin epoxidase, chloroplastic	18,633
SORBI_3006G097200	SNP_6_46696897	0.199	3.95E-15	1.205E-09	Translucence	BLINK	gamma-glutamylcyclotransferase 2-3	1,439
SORBI_3007G068300	SNP_7_7647653	0.392	5.31E-08	2.721E-03	Translucence	BLINK	polyphenol oxidase I, chloroplastic	30,159
SORBI_3007G068500	SNP_7_7647653	0.392	5.31E-08	2.721E-03	Translucence	BLINK	polyphenol oxidase II, chloroplastic	18,072
SORBI_3007G068700	SNP_7_7647653	0.392	5.31E-08	2.721E-03	Translucence	BLINK	polyphenol oxidase I, chloroplastic	17,865
SORBI_3003G111100	SNP_3_10030133	0.140	5.91E-07	1.802E-02	Translucence	BLINK	pathogen-related protein-like	21,133

Ensemble ID	SNPID	maf	P	FDR	Trait	Method	Annotated function	SNP distance from gene
SORBI_3003G111300	SNP_3_10030133	0.140	5.91E-07	1.802E-02	Translucence	BLINK	proteinase inhibitor PSI-1.2	3,856
SORBI_3007G068300	SNP_7_7647653	0.392	2.07E-09	3.151E-04	Translucence	FARMCPU	polyphenol oxidase I, chloroplastic	30,159
SORBI_3007G068500	SNP_7_7647653	0.392	2.07E-09	3.151E-04	Translucence	FARMCPU	polyphenol oxidase II, chloroplastic	18,072
SORBI_3007G068700	SNP_7_7647653	0.392	2.07E-09	3.151E-04	Translucence	FARMCPU	polyphenol oxidase I, chloroplastic	17,865
SORBI_3001G191200	SNP_1_16981727	0.357	3.35E-08	2.419E-03	Translucence	FARMCPU	dormancy-associated protein 1	24,337
SORBI_3006G096500	SNP_6_46696897	0.199	1.39E-07	7.045E-03	Translucence	FARMCPU	glutamate decarboxylase 2	39,684
SORBI_3004G201100	SNP_4_55263055	0.384	9.33E-07	2.021E-02	Translucence	FARMCPU	flavonoid 3'-monooxygenase	within Gene
SORBI_3004G200800	SNP_4_55263055	0.384	9.33E-07	2.021E-02	Translucence	FARMCPU	flavonoid 3'-monooxygenase	41,956
SORBI_3004G200900	SNP_4_55263055	0.384	9.33E-07	2.021E-02	Translucence	FARMCPU	flavonoid 3'-monooxygenase	29,472
SORBI_3006G096500	SNP_6_46696897	0.199	3.37E-07	7.378E-03	Translucence	MLM	glutamate decarboxylase 2	39,684
SORBI_3006G096500	SNP_6_46657114	0.121	7.28E-07	1.233E-02	Translucence	MLM	glutamate decarboxylase 2	99
SORBI_3007G151400	SNP_7_58345307	0.279	1.43E-06	4.838E-02	Translucence	MLM	cytokinin dehydrogenase 11	within Gene

Table 2-11 Genes linked to SNPs associated with bioclimate variables identified through genome-environment Association Analysis.

Ensemble ID	SNPID	maf	P	FDR	Trait	Method	Annotated function	SNP Distance from gene (bp)
SORBI_3001G187000	SNP_1_16432372	0.228	4.29E-13	1.307E-07	Annual Precipitation	BLINK	4-coumarate--CoA ligase-like 1	within Gene
SORBI_3009G228600	SNP_9_56994935	0.414	4.79E-08	3.651E-03	Annual Precipitation	BLINK	cytochrome c oxidase assembly protein COX19	48,450
SORBI_3001G187000	SNP_1_16432372	0.228	9.1E-09	2.774E-03	Annual Precipitation	FARMCPU	4-coumarate--CoA ligase-like 1	within Gene
SORBI_3009G198200	SNP_9_54857300	0.241	2.7E-09	8.224E-04	Precipitation of coldest Quarter	FARMCPU	28 kDa heat- and acid-stable phosphoprotein	12,863
SORBI_3002G061800	SNP_2_5947146	0.373	2.24E-07	2.073E-02	Precipitation of June	FARMCPU	protein argonaute MEL1	35,845
SORBI_3003G268700	SNP_3_60581629	0.241	3.14E-09	4.997E-04	Precipitation of October	FARMCPU	mitogen-activated protein kinase kinase kinase A-like	49,750
SORBI_3003G268800	SNP_3_60581629	0.241	3.14E-09	4.997E-04	Precipitation of October	FARMCPU	mitogen-activated protein kinase kinase kinase YODA	14,413
SORBI_3003G268900	SNP_3_60581629	0.241	3.14E-09	4.997E-04	Precipitation of October	FARMCPU	mitogen-activated protein kinase kinase kinase 2	8,931
SORBI_3003G269000	SNP_3_60581629	0.241	3.14E-09	4.997E-04	Precipitation of October	FARMCPU	mitogen-activated protein kinase kinase kinase 3	within Gene
SORBI_3005G015450	SNP_5_1400244	0.339	1.37E-07	1.389E-02	Precipitation of October	FARMCPU	cytochrome P450 714C3-like	536
SORBI_3005G015600	SNP_5_1400244	0.339	1.37E-07	1.389E-02	Precipitation of October	FARMCPU	cytochrome P450 714C2	7,763
SORBI_3009G198200	SNP_9_54857300	0.241	2.4E-07	2.054E-02	Precipitation of September	FARMCPU	28 kDa heat- and acid-stable phosphoprotein	12,863

Ensemble ID	SNPID	maf	P	FDR	Trait	Method	Annotated function	SNP Distance from gene (bp)
SORBI_3010G080400	SNP_10_6805250	0.230	2.7E-07	2.054E-02	Precipitation of September	FARMCPU	ethylene-responsive transcription factor ERF014	45,091
SORBI_3002G427900	SNP_2_77468713	0.074	7.3E-07	3.438E-02	Precipitation of September	FARMCPU	senescence associated gene 20	36,429
SORBI_3005G107800	SNP_5_20464774	0.101	1.4E-06	4.751E-02	Precipitation of September	FARMCPU	chalcone synthase 1	7,036

Chapter 3 - Germplasm sampling strategy affects the performance of genomic prediction: A case of Ethiopian Sorghum Landraces

Abstract

Germplasm screening is a vital avenue to utilize naturally available genetic variation for crop improvement. However, the inherent population structure of germplasm collections and their colossal size make screening for promising genotypes a daunting task. In this study, the potential use of genomic prediction was assessed using 304,802 SNPs and more than 1400 diverse Ethiopian germplasm was evaluated in multi-environment trials. First, the genomic prediction (GP) accuracy was assessed using two models – genomic best linear unbiased prediction (gBLUP) and ridge regression BLUP (rrBLUP) on the BLUP values of phenological and grain attributes. We computed the validation accuracy of the models utilizing training-set sizes ranging from 25 to 500 genotypes. The result showed that both models had comparable validation accuracies for all traits and training sizes. Generally, increasing training size increased validation accuracy. Days to flowering had the highest (0.70) validation accuracy, followed by plant height (0.66) and days to maturity (0.61). Hundred kernel weight (HKW) had moderate (0.49) prediction accuracy, while grain yield (0.39) and grain protein content (0.34) had the lowest validation accuracies at the training-set size of 500. Second, the effect of FIGS sampling method on the overall accuracy of GP was assessed. The FIGS approach utilizes landrace origin information to narrow down germplasm to a smaller target-trait enriched population. In this study, seed mass was used as a target trait and a proxy for assessing the FIGS effect on the general population parameters and GP accuracy. The FIGS sampling approach reduced the average pairwise distance between individuals in the established reference population, increased average genome-wide LD, and changed race composition relative to the base population. The GP accuracy implemented on FIGS-derived reference population was low relative to the GP on random reference populations. A modified GP which included germplasm from contrasting environments improved GP and selection differential. Overall, the result showed that GP can be implemented on a diverse germplasm population and offers an outline for the future design of training populations utilizing the FIGS approach.

Introduction

The genetic gain in crop improvement has either slowed or plateaued in many crops partly due to overexploited genetic diversity within elite lines (Grassini et al., 2013; Rakshit et al., 2014). This is particularly concerning given that climate change and population pressure are poised to strain food availability by reducing productivity and increasing demand for food. Hence, developing efficient methods to maximize functional genetic variation and exploit them in breeding programs is of paramount importance to improving genetic gain (Fonseca et al., 2021b). Cognizant of this, breeders, in addition to improving target trait, always strive to expand genetic variability through enriching the breeding population with fresh germplasm materials (Wang et al., 2017). However, identifying new lines and alleles from a large pool of uncharacterized germplasm is usually time-consuming and often ends up with little or no success.

One tool gaining traction to characterize and select promising germplasm is genomic prediction (GP) (Yu et al., 2016; Dzievit et al., 2021; Fonseca et al., 2021a). It leverages the recent advances in genotyping technologies, statistical modeling, and computing power. GP is based on training a model using a subset of accessions, training population, and later expanding the trained model to predict the breeding value of genetically related un-phenotyped test-set (Meuwissen et al., 2001). Unlike the QTL-based marker-assisted selection which only uses specific large effect markers for selection, GP, through simultaneous estimation of genome-wide additive effects, predicts the genomic estimated breeding values (GEBVs). It can also predict the total genotypic value by modeling additive, dominance, and epistatic genetic effects. Numerous models have been proposed for estimating genotypic values, which can be broadly categorized into parametric and non-parametric models. The parametric models utilize assumptions about population-based parameters and include the BLUP models: rrBLUP (Whittaker et al., 2000; Meuwissen et al., 2001), G-BLUP (VanRaden, 2008), Least Absolute Shrinkage and Selection Operator LASSO (Usai et al., 2009), and the Bayesian alphabet models BayesA, BayesB, BayesC (Meuwissen et al., 2001). The non-parametric models, however, do not rely on population assumptions and some of these include Reproducing Kernel Hilbert Spaces regression (RKHS) (Gianola et al., 2006), machine learning methods including support vector regression (Moser et al., 2009) and random forests (González-Recio and Forni, 2011). The G-BLUP and rr-BLUP are based on the infinitesimal additive model and estimate mainly additive genetic effects. Non-parametric and

sometimes parametric models, using appropriate kernel function, can model total genotypic value by modeling epistatic and dominance interactions (Endelman, 2011; González-Camacho et al., 2012; de Los Campos et al., 2013). A comparison of the different, more complex models such as Bayesian alphabet models and LASSO with respect to GBLUP and RR-BLUP showed mixed result. gBLUP and rrBLUP performed comparably primarily when empirical data was used (Maulana et al., 2021; Ganesamurthy et al., 2022; Meher et al., 2022). In studies that involved germplasm collection, comparable results had been reported with rrBLUP and gBLUP producing similar accuracy levels with Bayesian models (Yu et al., 2016).

GP has become an essential breeding tool mainly for developing elite breeding lines, while its utilization for exploring landrace germplasm is minimal but slowly picking up (Crossa et al., 2016; Yu et al., 2016; Muleta et al., 2017). Factors that generally affect the performance of GP for advanced lines, such as heritability, the genetic architecture of the trait, and population structure (Fonseca et al., 2021a) would similarly affect GP for germplasm screening. The major constraint, specifically for germplasm screening, emanates from the loss of prediction accuracy for collections inherently possessing a strong population structure. Because the germplasm collections come from diverse agroecology, reproductive isolation and adaptive evolution produce genetically distinct subpopulations with extensive LD decay. The predictive accuracy of GS models declines when the training and validation populations are genetically distant and when the LD pattern between the training and the validation population changes (Clark et al., 2012; Lorenz and Smith, 2015). Such conditions are expected to be common in landrace collection. The ideal condition for higher accuracy of GP is where both the training and test sets share the same casual mutations (Olatoye et al., 2020). Different strategies have been suggested to circumvent the negative effect of population structure. These include stratified analysis (Hayes et al., 2009; Olson et al., 2012) and allele-cluster interaction to model heterogenous allele effect across breeds (de Los Campos et al., 2015) with a slight improvement in the overall accuracy of genomic prediction (de Los Campos et al., 2015).

Sorghum is one of the major crops with large germplasm collections at different gene banks worldwide. It has magnificent resilience to marginal environments where moisture conditions are unpredictable. This specific attribute of sorghum makes it a crop of choice to combat climate change. Due to its broad adaptation to a range of agroecology, it is expected to harbor a wealth of

genetic resources waiting to be exploited. More than 165,000 germplasms are stored globally in different locations. Such an extensive landrace collection is difficult to evaluate because of logistical, resource, and technical constraints. GS offers the tool to minimize the sample needed to be assessed phenotypically and can help predict the genetic value of un-phenotyped germplasm. However, for GP and GS to work, the accessions need to be genotyped, and even that is still impractical given the sheer size of germplasm in gene banks. One strategy to maximize exploitation of such germplasm is to utilize the center of origin and diversity of landraces as a starting point. Countries like Ethiopia that harbors the greatest diversity of the crop may serve as focal geography for such study. Ethiopian sorghum germplasm has been the source of many vital traits under discovery and utilization worldwide (Singh and Axtell, 1973; Haussmann et al., 2002; Rhodes et al., 2017; Nida et al., 2019; Muleta et al., 2021). The Ethiopian materials comprise all major botanical races and their intermediates (Doggett, 1988), with durra, bicolor, and caudatum and their mixed races being dominant. The majority of sorghum germplasm collections maintained in different parts of the world constitute a great deal of the Ethiopian germplasm. The local gene bank in Ethiopia, the Ethiopian Biodiversity Institute, maintains some 9772 sorghum germplasm accessions collected across the country, about 9760 accessions of Ethiopian origin are represented in international gene banks (CGIAR, 2007).

GS and GP may be applied both for genetic improvement of the crop and to maintain diversity in germplasm conservation endeavors. In cases where the objective is to preserve the diversity of the population within the sampled core collection, a representative sampling from across agroecology of origin is implemented (Upadhyaya and Ortiz, 2001). However, rare and beneficial alleles may not be identified using this approach (Street et al., 2016). Whereas in situation where germplasm is mined for trait mining, the FIGS approach may be used to enrich the selected landraces with favorable alleles of the target trait. The FIGS approach works with the premise that selection increases the frequency of favorable alleles. A *priori* relationship between environmental parameters of origin and a trait of interest can be used to predict potential landraces with the target attribute. The FIGS approach had been successful in identifying rare variants for different adaptive traits, including sunn pest (Bouhssini et al., 2009), and Russian wheat-aphid (Bouhssini et al., 2011) in wheat, and drought-adaptation in faba bean (Khazaei et al., 2013).

The size of the reference population constituted through the FIGS approach is variable across the literature. In published reports - 1125 from 17778 (El Bouhssini et al., 2011); 500 from 3738 (Endresen et al., 2012); 87 from 4576 (Dadu et al., 2019). The size may depend on a variety of factors including but not limited to the type of species, the base-population size, the size of collection from a favorable environment, and the relative ease of the trait to phenotype. As the FIGS selection is made using environment-only information, the approach is not directly selecting genotype but environments. As a result, the approach does not have the power to distinguish genotypes originating from similar environments. Such concern is especially critical in centers of diversity such as Ethiopia, where diverse germplasm exists in proximity. In such cases, these diverse sets of landraces tend to possess common environmental parameters and sampling only a few based solely on environmental data may risk missing beneficial variants hiding in another environment. However, GP can complement the FIGS approach by letting the pipeline to sample first a larger set from the germplasm collection. Later, a smaller training population with manageable size can be sampled to conduct thorough phenotyping and to train genomic prediction. Such an approach may increase the search space and improve chance of identifying potential landraces.

Combining both approaches may bring complementarity where the FIGS approach removes the garbage and paves the way to 'identify the needles in a haystack' (Shim et al., 2021), while GS opens the opportunity to evaluate a more extensive set with the potential to identify promising landraces with beneficial sets of alleles. Nevertheless, coupling FIGS and GP may have an impact on the overall accuracy of the pipeline as the FIGS sampling strategy to form the reference population can affect population parameters. The FIGS sampling may return subpopulations with narrower genetic base because of the likelihood of selecting genetically related individuals from similar environments. Since germplasm collections come from landraces distributed across agroecology, population structure is likely to present in such collections. Studies have showed that population structure affects the overall accuracy of genomic prediction (Muleta et al., 2017; Sapkota et al., 2020). In this study, we first evaluated the overall performance of GS in Ethiopian sorghum core collection using phenological traits and physicochemical grain attributes. Second, we assessed the effect of training size on the accuracy of GP. Third, we evaluated the impact of the FIGS sampling approach on the accuracy of GP and the overall performance of the FIGS-GS pipeline.

Materials and methods

Plant materials and study sites

A population panel comprising 2010 accessions including landraces, improved varieties, and inbred lines was used in this study. The collections were assembled as part of a collaborative effort supported by the United States Agency for International Development (USAID) through the sorghum and millet innovation lab (SMIL) in partnership with US and Ethiopian Universities, Ethiopian Institute of Agricultural Research (EIAR) and Ethiopian Biodiversity Institute (EBI). The testing sites are outlined in (Tessema et al., 2019) and briefly summarized as follows. The Field experiments were laid out at Haramaya University, Arsi-Negele, Bako, Pawe, and Meiso sites, in different regions of Ethiopia. Each accession was planted in a non-replicated 3m long single rows, at a spacing of 75 cm between rows, and a 20 cm between plants. The experiments were planted at the regular planting time for the crop, from mid-April to mid-May. At planting, phosphorus was applied as di-ammonium phosphate (DAP) at the rate of 46 kg ha^{-1} P_2O_5 and 18 kg ha^{-1} N. Additional nitrogen fertilizer was applied in the form of urea at the rate of 46 kg N ha^{-1} when the crop was around knee height. Fields were regularly supervised by resident technicians at each station and plots were kept free of weeds by manual weeding. Data were collected on a number of traits including emergence, flowering, plant height, maturity, and leaf and panicle diseases. After harvest, more data were collected on grain quality, yield components, and protein content. Only those data applicable to this study are further described herein. Days to flowering was recorded as the number of days from planting until 50% of the plants in a plot reached half bloom; while days to maturity as the number of days taken for grains in the middle section of the panicle reached the black layer stage. Plant height was measured as the average length of a mature plant measured from the base. Hundred kernel weight (HKW) was measured as the weight of 100 kernels and adjusted to 12% moisture content. Yield per panicle was estimated as the average weight of grains from five random panicles collected from each plot. Grain protein content was estimated using PertenIM-9500 (PerkinElmer), and later, the spectral data was calibrated for protein content determination at USDA, Manhattan, Kansas, USA. The geographic description of each trial location is shown in Table 3-6.

Genotypic data

A portion of the collection, 1628 accessions were genotyped by the genotyping by sequencing (GBS) platform. The procedure for DNA extraction, library preparation, and sequencing were as described in Tessema et al. (2019). Briefly, DNA was isolated using the CTAB procedure from the landraces (Mace et al., 2003). The GBS genotyping was carried out at the University of Wisconsin Biotechnology Center. The TASSEL Version 5 (Bradbury et al., 2007), GBSV2 pipeline (Glaubitz et al., 2014) was utilized to process the raw sequence files, aligning to the *Sorghum bicolor* reference genome version 3.1.1 from Phytozome (McCormick et al., 2018). A "very-sensitive" parameter of Bowtie2 (Langmead and Salzberg, 2012) was used to call single nucleotide polymorphism (SNPs). The GBS pipeline produced a total of 397,313 SNPs across the landraces. These were further filtered for <12.5% Heterozygosity, biallelic sites, a maximum of 20% missing data (Dzievit et al., 2021), and a minimum of 1% minor allele frequency using VCFtools - 0.1.17 (Danecek et al., 2011), finally retaining 304,802 SNPs. Imputation for missed loci was performed separately for each chromosome using Beagle4.1

Bioclimatic parameters

The passport data of accessions included district, zone, and administrative regions. Bioclimatic data was obtained from the WorldClim Version2 database (Fick and Hijmans, 2017). Administrative boundaries were retrieved from shape files archived in the Open Africa Database (https://africaopendata.org/). R code Package- Raster (Hijmans et al., 2017) was used to obtain the mean value representing the respective district for 19 climate variables. Bioclimatic data was extracted for 1258 of the landraces with genotypic data.

Plant parameters

The BLUPs values of all accessions was used for all traits to account for the environment effect using a mixed linear model implemented in the R package *lme4* (Bates et al., 2015). The BLUPS were estimated as follows:

$$y_{ij} = \mu + G_i + L_j + \varepsilon_{ij}$$

Where y_{ij} is the observed phenotypic value of i^{th} accession in the j^{th} environment, μ is the overall mean, G_i is the random genotypic effect for the i^{th} accession, L_j is the fixed environmental effect of the j^{th} envirionment, and ε_{ij} is the residual.

Variance components and heritability

Components of variance were computed using the *lmer* package treating genotype effect as random and environment effect as fixed. Broad-sense heritability on the entry-mean basis was roughly estimated using the equation:

$$H = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_{e/l}^{2\,2}}$$

Where $\sigma^2_{g\ \text{and}}$ σ_e^2 are estimated genetic error variances, respectively. *l* is the number of environments.

Principal component analysis, linkage disequilibrium and genetic distance

Population structure and landraces' relatedness were established using principal component analysis and kinship matrices. We used the Tassel (Bradbury et al., 2007) command-line version to run principal component analysis, while the R package-GAPIT (Lipka et al., 2012) was used to compute the Kinship matrix. We manually assigned botanical race information from our unpublished work following the procedure outlined in Harlan and de Wet (1972; IBPGR and ICRISAT (1993). The supervised option of the Admixture option (Alexander et al., 2015) was used to assign the majority of the unassigned landraces using botanical race data obtained from Wang et al. (2020a). Linkage disequilibrium for the different sets of populations was separately characterized by PLINK (Purcell et al., 2007) where r^2 was computed using 100 SNPs within 100 Mb window.

Effect of prediction method and population size on accuracy of genomic prediction

The study first evaluated the genomic prediction accuracy of different plant traits using rrBLUP and GBLUP models implemented in the R package *rrBLUP* (Endelman, 2011). The rrBLUP is explained as follows:

$$y_i = \mu + \sum_{k=1}^{p} X_{ik}\,\beta_k + \varepsilon_i$$

Where y_i is the observed phenotype of the i^{th} individual, μ the mean, X*ik* is the genotype matrix for biallelic single nucleotide polymorphisms of i^{th} individual and k^{th} marker. β_k is additive random effect of the k^{th} marker, k~N $(0,\sigma_g^2)$ and ε_i is the residual error ~N $(0,\sigma_e^2)$. GEBVs for individuals were calculated as the sum-total of marker effects.

We used the GBLUP method introduced by Habier et al. (2007) and VanRaden (2008) and implemented using rr-BLUP software package where:

$$y = 1\mu + Z\gamma + e$$

Where y is phenotype vector, 1 is a vector of 1s, μ is grand mean, Z is the incidence matrix for breeding values, e is the residual error, $e \sim N(0, \sigma_e^2)$, and γ is a vector of breeding values, ~N $(0, G\sigma_g^2)$. σ_g^2 is genetic variance, and G is calculated as follows:

$$G = \frac{WW'}{2\sum p_i(1-p_i)}$$

Where W is computed from a $n \times m$ marker matrix coded as (0,1,2) by adding $-2p_i$, where p_i is the allelic proportion of one of the alleles in the population.

At this early stage of breeding, the interest is mainly on additive effects. Both rrBLUP and GBLUP are computationally faster models capable of modeling additive effects. Their performance is as good as the more complex models when used to screen a wide range of traits for GS (Bhering et al., 2015; Yu et al., 2016).

Performance of GP across multiple traits

Of the 2010 accessions included in the trial, we used only those for which genotypic data available (1628 accessions). Training sizes of 25, 50, 100, 150, 200, 250, 400, and 500 accessions were used to evaluate predictive accuracy for both rr-BLUP and G-BLUP methods. A thousand iterations were made for each training size. Mean Pearson correlation coefficient r between GEBV and observed values was estimated for the training population. The model developed using the

specified training population estimates GEBVs for the remaining validation population. The mean correlation of the observed phenotypic values and the test set's respective GEBVs was considered validation accuracy. Unless stated, accuracy in the text refers to validation accuracy. The plateau for prediction accuracy of increasing training population size was determined as the tilt below the threshold of 0.0001 of the rate of improvement in accuracy per unit training size increment.

Effect of FIGS sampling approach to select larger seed mass on the overall Performance of GP

The FIGS approach works mainly for adaptive traits where a relationship between trait of interest and landrace origin can be made. Such traits include biotic, and biotic stress tolerance, where environmental cues can be associated with plant attributes. For this study, we retained HKW as a target trait for the FIGS approach. In sorghum, seed mass is one of the important adaptation-related traits associated with precipitation gradient (Wang et al., 2020a). Different studies suggest that landraces from drier areas possess larger seed masses (Stromberg and Boudell, 2013; Wang et al., 2020a). For the FIGS sampling method, the approach by (Khazaei et al., 2013) was adopted with modification. Since seed mass had been associated with precipitation gradient, bioclimatic variables associated with magnitude of precipitation were used. The computed Euclidean distance was computed for the environments and distance parameter for hierarchical clustering was used. The clusters were sorted based on the annual precipitation from the driest to the wettest. Individuals were selected from cluster with lowest annual precipitation and continued to the next cluster until the size of the reference population, which was set to 700, was met.

The pipeline which incorporates the GP in the FIGS approach employs two sampling stages, as outlined in Figure 3-1. The first sampling step is where we mimicked the scenario where large gene banks are sampled to a manageable size of germplasm collection, which we coin henceforth as the reference population. The reference population in some literature is synonymous with the training set (Zhu et al., 2021). However, in this study, the reference population is the working population which comprises both the training population and the test sets (Yu et al., 2016). We used three methods at this sampling stage: The FIGS approach, here termed as FIGS_Dry approach, where we followed the procedure outlined before. A modified version of the FIGS we called FIGS staggered where we sampled 80% and 20% of the individuals from the driest and

wettest clusters, respectively. The other sampling method used to establish reference populations was the random sampling method. We sampled 50 random and independent reference populations using this sampling method (Figure 3-11).

The second sampling stage draws individuals from the reference population to set up training populations for GP (Figure 3-1). For this stage, we used random sampling from the reference populations constituted using FIGS_Dry, FIGS_Staggered and random methods. Additionally, representative sampling of the reference populations formed by FIGS_Dry and FIGS_Staggered approaches was made based on K-means clustering on PCs of the genomic data. For simplicity, we will refer FIGS_Dry approach as FIGS approach and FIGS_staggered approach as staggard approach from this point forward. For each reference population, one hundred independent training sets were determined. Training and validation accuracies were computed for each iteration of the training population. For each reference population, average of these iterations is reported. We also compared FIGS-GP and staggered-GP pipelines in terms of the average of observed HKW values of top landraces with 5% and 10% GEBVs.

We tested the statistical significance of whether proportions changed, enriched, or shrank when we drew using FIGS approaches from the base population using two-tail, right tail, or left tail binomial tests implemented in R (R Core Team, 2020), respectively. Significance difference in means was tested using a two-tailed t-test using base R statistical software (R Core Team, 2020).

Result

Landraces performance, variance components, and heritability

The grain and phenological data of the landraces across environments is shown in Table 3-1. Mean days to flowering ranged from 66 to 159, which is typical for tropical landraces, and took 125 to 214 days to mature. Mean plant height (PH) ranged from 98 to about 467 cm. HKW and grain protein content ranged from 0.5 to 4.6 g and 7.1 to 15.5%, respectively. Yield per plant also went from 5 to 174 g per panicle, with a mean of 55.7 g (Table 4-1). The heritability of the traits was also variable, where plant height had the highest heritability (90.3%) followed by days to flowering (88.4%) and days to maturity (68.1%). Grain protein content (64.2%) and yield per panicle (51.5%) had the least heritability.

Training size influences genomic prediction

Using the entire 2010 population as a reference, we estimated the validation accuracy of GBLUP and rrBLUP (Table 3-2 and Table 3-3, respectively) using models established by training 500 individuals. The training population was sampled randomly for both models. Days to flowering (r_{rrBLUP} =0.698, r_{GBLUP}=0.699) followed by plant height (r_{rrBLUP} =0.661, r_{GBLUP}=0.665) and had high prediction accuracy of the validation population compared to the grain attributes HKW ($r_{-rrBLUP}$ =0.489, r_{-GBLUP}=0.488), grain protein content (r_{rrBLUP} =0.333, r_{GBLUP}=0.333) and yield per panicle (r_{rrBLUP} =0.389, r_{-GBLUP}=0.396). A comparison of the validation prediction accuracy relative to the estimated broad-sense heritability computed as a ratio of (Table 3-1) revealed that days to maturity had the highest ratio (0.90), while grain protein content had the least (0.61)

Both models fit by rrBLUP, and gBLUP yielded comparable results for all the training sizes (Table 3-2 and Table 3-3). In all traits, increasing population size improved the prediction accuracy of the validation population. The most significant improvement (83%) was obtained for yield per panicle, followed by days to maturity (56.5%) and grain protein content (48%), while HKW did not show much improvement (only 23%). While we observed improvement with an increase in training size, the rate of improvement per a unit increment of training size kept declining. The prediction accuracy improvement per unit training size increase approached below a training size around 300 < 0.0001 for most traits.

FIGS sampling approach to sample landraces with larger HKW

Assuming that genotype and phenotype information are not available, we selected the reference population using only the passport data (Figure 3-1). The site of origin was used to extract bioclimatic variables. Using the clustering approach of bioclimatic precipitation variables, we obtained nine optimal clusters determined by using the hierarchical clustering approach (Figure 2A). A reference population size of 700 landraces was established by drawing from each cluster sorted by their mean annual precipitation (Figure 3-2 B & C). The mean HKW of landraces grouped in the respective clusters was correlated with the mean annual precipitation of clusters (r= -0.829, P=0.0057) (Figure 3-2 and Figure 3-2D). Botanical race membership also differentiated

the population into different grain weight classes (Figure 3-3 A). Selection based on precipitation gradient yielded a significant difference in HKW ($P \leq 0.0001$) (Figure 3-2 D) and also caused compositional change of landraces (Table 3-4). Durra had a 14% ($P=0.0009$) increase in the proportion of the selected reference materials, while caudatum shrank by more than 40% ($P < 0.0001$). The FIGS approach significantly enriched HKW relative to the whole panel and to the staggered approach (Figure 3-3B).

Principal component analysis reveals population structure within the whole panel where botanical races are clustered distinctively (Figure 3-4). LD analysis showed that LD decayed to half in the whole panel, FIGS, and staggered reference population within 2.5 kbp, 3kbp, and 3kbp, respectively. As expected, the average pairwise distance computed from IBS showed that the whole panel is composed of genetically distant landraces (0.309) followed by staggered reference population (0.306). The FIGS population was constituted relatively by more related individuals (0.303) (Table 3-5).

Origin-based sampling impacted overall validation prediction accuracies

Training prediction accuracies were higher than validation prediction accuracies (Figure 3-5). The median training accuracy for the randomly established populations was higher than all origin-based approaches across all the training sizes. The training accuracies of FIGS at 100 and 400 training sizes were higher than the staggered training accuracy, while at 200 training population size, the staggered approach outperformed the FIGS approaches. Like the training accuracy, median validation accuracies from the random reference populations had higher accuracy than the origin-based methods. The staggered approach had higher validation accuracy than the FIGS approach for all the training sizes (Figure 3-5).

To evaluate the GP-assisted performance of the FIGS approach in identifying the best individuals from the whole panel, we considered the mean observed HKW of the top 5% and 10% of individuals selected based on GEBVs (Figure 3-6). As expected, the smaller proportion (5%) had higher mean performance than the higher proportion (10%). However, a similar pattern of ranking for the selection approaches was observed for both ratios. We compared relative performance with the mean distribution of top GEBV individuals from the reference populations.

The FIGS approach at 25 and 50 training sizes trailed far to the left tail of the random reference-population mean distribution.

In contrast, the staggered approach performed better and stood around the middle of the distribution. However, the FIGS approach shot up at 100 and 150 training sizes and marginally outperformed the staggered approach. In larger training populations starting from size 200, the FIGS approach retracted back while the staggered approach consistently improved for all training sizes (Figure 3-6).

Discussion

In light of the anticipated increase in food demand and a production challenge posed by climate change, breeders need to adopt new methods, tools, and technologies to improve the attributes of food and feed crops to satisfy human needs. Climate change is predicted to bring about new challenges to crop production, such as increased disease and pest prevalence and unpredictable weather (drought, heat) that may alter crop adaptation and productivity. Coping with such changes may require, among others, exploring unutilized genetic resources stored in gene banks. The major hurdle to properly accentuate on this approach is the sheer size of germplasm collection stored in gene banks and the lack of clue where the desired allelic variants may be hiding. This task of chasing a slim probability of identifying a genotype of interest from the large gene pool is appropriately compared with trying to find a needle in a haystack (Shim et al., 2021). The advent of GS, complemented by the recent advances in genotyping, phenotyping, and computational capabilities, may provide a useful tool to facilitate the mining of germplasm resources and increases the chance of success in identifying the genotypes/genes of interest.

In this study, we first evaluated the potential use of GS on diverse Ethiopian germplasm using a range of parameters. Among the phenological traits, our prediction accuracy for plant height of 0.65 was comparable to other reports on sorghum (Yu et al., 2016; Habyarimana et al., 2020). Grain weight, highly influenced by population structure, had moderate prediction accuracy (0.49) across environments. Sapkota et al. (2020) reported a slightly higher (0.65) prediction accuracy for grain weight evaluated across botanical races, while they reported smaller accuracy (0.31) for the within race prediction. GPC prediction accuracy is low for both rrBLUP and GBLUP (~0.33). Studies about the genomic prediction for the GPC in sorghum are scarce. In wheat, a

moderate level of accuracy (0.41) was reported (Huang et al., 2016). Higher single environment prediction accuracy (0.50 to 0.69) was reported using the NAM population (Sandhu et al., 2021), while combined multi-year prediction accuracy declined to 0.3 to 0.43. They attributed the low multi-environment prediction accuracy to genotype-by-environment interaction, where the importance of loci determining GPC becomes different for different environments. Similarly, the prediction accuracy for grain yield per panicle (0.33) was also low but comparable to previously reported values by Velazco et al. (2019) and Hunt et al. (2018).

Training population size is an important factor in determining prediction accuracy. For all traits, we observed an improvement in genomic prediction accuracy with an increase in population size. Similar results were also reported in maize (Zhang et al., 2017), wheat (Muleta et al., 2017), and rice (Berro et al., 2019). With increasing population size, the marker effects could be more accurately estimated and may result in better accuracy (Muleta et al., 2017). In the current study, while training size improved prediction accuracy, the rate of improvement (relative change of accuracy per a unit increase in training size) with further increase in the size of training set declined rapidly and reached our threshold < 0.0001 around training size 300.

Reference populations sampled from a larger germplasm collection need to be as large as possible to minimize the risk of missing promising, and as small as possible to reduce the resource burden. Different strategies are utilized to narrow down larger germplasm collection in gene banks to make up the reference population. Generally, traits related to crop adaptation are correlated with agroecology of the adaptation region. Approaches like FIGS utilize this relationship to narrow down potential genotypes for adaptation-related characteristics. For this study, we selected grain weight, as associated with adaptation to dryland environments. Even though the mechanism behind it is not yet established, it has been hypothesized that larger seed weight offers resources in the initial stage of emergence and crop establishment, playing a central role in survival under dry conditions (Stromberg and Boudell, 2013; Wang et al., 2020a). In sorghum, grain weight is correlated with precipitation variables which can easily be cross-referenced using the place of origin. Additionally, grain weight is highly associated with population structure, affecting the overall accuracy of genomic prediction. As a result, it is a good proxy parameter for evaluating the effect of using origin-based reference population establishment on the performance of GS, which can be extended to other adaptation-related traits.

As shown in Figure 3-3, the reference population selected using the FIGS approach had a mean seed mass larger than the base population. Since genotypes adapted to similar environments also tend to be in relative physical proximity to each other or have some degree of genetic similarity due to genetic parallelism (Passarella et al., 2008), we expected the FIGS approach to yield more related genotypes. As expected, the average pairwise genetic distance of individuals from the FIGS reference population was the smallest compared to the whole panel and the staggered reference population (Table 3-5). Moreover, the large-seeded durra types that dominate the country's drier regions were enriched in the population, contributing to the higher mean seed weight, while medium-sized caudatum types prevalent in humid areas were underrepresented.

In addition to its association with geographic origin, grain weight was also highly associated with population structure, further confirming its potential use as a proxy parameter for evaluating the effect of using origin-based reference population establishment on the performance of GS, which can be extended to other adaptation-related traits. As the population composition changed, the LD pattern also changed where the FIGS reference population had, on average, more extended LD (59kbp) than the whole panel (40 kbp). Similar changes in composition and the consequent changes in population parameters had been shown to impact overall genomic prediction accuracies. In sorghum, a study on the effect of population composition on validation accuracy of genomic prediction showed that models trained in relatively homogeneous populations tend to fail when used to predict a more heterogeneous population (Sapkota et al., 2020). Similarly, in wheat, genomic prediction using two distantly related groups where one was used as a training and the other as a test population had inferior prediction results to the GP model utilizing the mixture of the two groups as a training population (Muleta et al., 2017). In rice and wheat, the genetic relationship between training and validation populations had paramount importance in shaping model accuracy on the test set (Berro et al., 2019). In this study, the over-representation of the durra race in the FIGS reference population, which is already significant in the base collection, may have reduced the overall stratification of the FIGS reference population. As a result, marker effects for the underrepresented group/s may not be estimated and as a result become source of bias. In a GP study involving cattle, validating a model on a validation set dominated by breeds that were the minority in the training set had lower validation accuracy. The accuracy was improved by incorporating multiple breeds in the training set (Olson et al., 2012).

In contrast, the staggered approach sampled 80% of individuals targeting the driest environment and the rest 20% from the wettest environments. This sampling procedure was aimed at capturing more diverse set of genotypes in the population while reasonably maintaining the population enriched for the trait of interest. The staggered reference population had higher validation accuracies than the FIGS population for all training sizes. The result conforms with (Isidro et al., 2015), who reported the stratified sampling under a strong population structure had yielded higher GP accuracy. The inclusion of environments from the extreme end of the precipitation gradient might have included genetic groups which were underrepresented in the FIGS reference population. Increasing the share of an underrepresented group in the reference population might have increased the chance of the minority groups occurring in the training set and better model marker effects. The incorporation of genotypes from the extreme environments might have also increased the overall variance of the trait in the training populations sampled. Larger phenotypic variance in a training population with a strong population structure had been associated with improvement in prediction accuracy (Isidro et al., 2015)

At last, we evaluated the mean HKW performance of the top 5 and 10% of individuals selected based on GEBVs. Generally, the top GEBV individuals from the FIGS reference population had the least selection differential than the reference populations developed through the staggered or random approach. However, for some training sizes, it outperformed the staggered method. The relatively smaller validation accuracy observed in the FIGS approach may have negatively impacted the accuracy of GEBV values, directly influencing the selection differential. However, the staggered approach improved as the training size increased, outperforming the FIGS approach in most training sizes. It is not clear why the FIGS approach performed well at 100 and 150 training sizes and retracted back with a training size above 150. Since the staggered approach comprised of individuals from both dry and wet environments, it may have increased the population's overall heterogeneity, opening the chance to assess the effect of genomic loci that would not otherwise be evaluated under the FIGS approach. The better estimation of marker effects may have improved the validation accuracy of test individuals.

Conclusion

The study was aimed at investigating the power of GS to exploit diverse germplasm collections for breeding purposes. The result showed that GS works for traits under consideration, but the model can be tweaked to enhance accuracy. The moderate validation accuracy observed in the smaller training set was improved by increasing training size, which can be optimized to provide decent prediction power. We also evaluated whether the FIGS approach, which ICARDA popularizes, is compatible with GS. Moreover, the FIGS approach was tested along with the GS to see the feasibility of the combined approach to targeting particular traits of interest in germplasm exploration. Since the FIGS approach is skewed in that it samples only individuals presumably possess known traits of interest, its accuracy in predicting the reference population parameter was only moderate. This was improved by using the staggered approach that allows the incorporation of individuals from either extreme for the trait of interest. The staggered approach improved validation accuracy and GEBV-based selection differential.

Reference

Alexander, D.H., S.S. Shringarpure, J. Novembre, and K. Lange. 2015. Admixture 1.3 software manual. Los Angeles: UCLA Human Genetics Software Distribution.

Bates, D., M. Mächler, B. Bolker, and S. Walker. 2015. Fitting Linear Mixed-Effects Models Using {lme4}. J Stat Softw 67(1): 1–48. doi: 10.18637/jss.v067.i01.

Berro, I., B. Lado, R.S. Nalin, M. Quincke, and L. Gutiérrez. 2019. Training population optimization for genomic selection. Plant Genome 12(3): 190028.

Bhering, L.L., V.S. Junqueira, L.A. Peixoto, C.D. Cruz, and B.G. Laviola. 2015. Comparison of methods used to identify superior individuals in genomic selection in plant breeding.

El Bouhssini, M., K. Street, A. Amri, M. Mackay, F.C. Ogbonnaya, et al. 2011. Sources of resistance in bread wheat to Russian wheat aphid (Diuraphis noxia) in Syria identified using the Focused Identification of Germplasm Strategy (FIGS). Plant Breeding 130(1): 96–97.

Bouhssini, M. El, K. Street, A. Joubi, Z. Ibrahim, and F. Rihawi. 2009. Sources of wheat resistance to Sunn pest, Eurygaster integriceps Puton, in Syria. Genet Resour Crop Evol 56(8): 1065–1069.

Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, et al. 2007. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23(19): 2633–2635.

CGIAR. 2007. Strategy for the Global Ex Situ Conservation of Sorghum Genetic Diversity.

Clark, S.A., J.M. Hickey, H.D. Daetwyler, and J.H.J. van der Werf. 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genetics Selection Evolution 44(1): 1–9.

Crossa, J., D. Jarqu'in, J. Franco, P. Pérez-Rodr'iguez, J. Burgueño, et al. 2016. Genomic prediction of gene bank wheat landraces. G3: Genes, Genomes, Genetics 6(7): 1819–1834.

Dadu, R.H.R., R. Ford, P. Sambasivam, K. Street, and D. Gupta. 2019. Identification of novel Ascochyta lentis resistance in a global lentil collection using a focused identification of germplasm strategy (FIGS). Australasian Plant Pathology 48(2): 101–113.

Danecek, P., A. Auton, G. Abecasis, C.A. Albers, E. Banks, et al. 2011. The variant call format and VCFtools. Bioinformatics 27(15): 2156–2158. doi: 10.1093/bioinformatics/btr330.

Doggett, H. 1988. Sorghum, 2nd edn. Tropical agricultural series.

Dzievit, M.J., T. Guo, X. Li, and J. Yu. 2021. Comprehensive analytical and empirical evaluation of genomic prediction across diverse accessions in maize. Plant Genome 14(3): e20160.

Endelman, J.B. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4(3).

Endresen, D.T.F., K. Street, M. Mackay, A. Bari, A. Amri, et al. 2012. Sources of resistance to stem rust (Ug99) in bread wheat and durum wheat identified using focused identification of germplasm strategy. Crop Sci 52(2): 764–773.

Fick, S.E., and R.J. Hijmans. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. International journal of climatology 37(12): 4302–4315.

Fonseca, J.M.O., P.E. Klein, J. Crossa, A. Pacheco, P. Perez-Rodriguez, et al. 2021a. Assessing combining abilities, genomic data, and genotype$\times$ environment interactions to predict hybrid grain sorghum performance. Plant Genome 14(3): e20127.

Fonseca, J.M.O., R. Perumal, P.E. Klein, R.R. Klein, and W.L. Rooney. 2021b. Combining abilities and elite germplasm enhancement across US public sorghum breeding programs. Crop Sci 61(6): 4098–4111.

Ganesamurthy, K., S. Das, R. Saraswathi, C. Gopalakrishnan, R. Gnanam, et al. 2022. Analysis of the Efficiency of Genomic Selection Models for Predicting Sheath Blight Resistance in Rice (Oryza sativa L.,). International Journal of Bio-Resource \& Stress Management 13(3).

Gianola, D., R.L. Fernando, and A. Stella. 2006. Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173(3): 1761–1776.

Glaubitz, J.C., T.M. Casstevens, F. Lu, J. Harriman, R.J. Elshire, et al. 2014. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS One 9(2): e90346.

González-Camacho, J.M., G. de Los Campos, P. Pérez, D. Gianola, J.E. Cairns, et al. 2012. Genome-enabled prediction of genetic values using radial basis function neural networks. Theoretical and Applied Genetics 125(4): 759–771.

González-Recio, O., and S. Forni. 2011. Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. Genetics Selection Evolution 43(1): 1–12.

Grassini, P., K.M. Eskridge, and K.G. Cassman. 2013. Distinguishing between yield advances and yield plateaus in historical crop production trends. Nat Commun 4(1): 1–11.

Habier, D., R.L. Fernando, and J. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177(4): 2389–2397.

Harlan, J.R., and J.M.J. de Wet. 1972. A simplified classification of cultivated sorghum 1. Crop Sci 12(2): 172–176.

Haussmann, B., V. Mahalakshmi, B. Reddy, N. Seetharama, C. Hash, et al. 2002. QTL mapping of stay-green in two sorghum recombinant inbred populations. Theoretical and Applied Genetics 106(1): 133–142.

Hayes, B.J., P.J. Bowman, A.C. Chamberlain, K. Verbyla, and M.E. Goddard. 2009. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genetics Selection Evolution 41(1): 1–9.

Huang, M., A. Cabrera, A. Hoffstetter, C. Griffey, D. Van Sanford, et al. 2016. Genomic selection for wheat traits and trait stability. Theoretical and Applied Genetics 129(9): 1697–1710.

Hunt, C.H., F.A. van Eeuwijk, E.S. Mace, B.J. Hayes, and D.R. Jordan. 2018. Development of genomic prediction in sorghum. Crop Sci 58(2): 690–700.

IBPGR, and ICRISAT. 1993. Descriptors for Sorghum (Sorghum bicolor (L) Moench). International Board for Plant Genetic Resources, Rome.

Isidro, J., J.-L. Jannink, D. Akdemir, J. Poland, N. Heslot, et al. 2015. Training set optimization under population structure in genomic selection. Theoretical and applied genetics 128(1): 145–158.

Khazaei, H., K. Street, A. Bari, M. Mackay, and F.L. Stoddard. 2013. The FIGS (Focused Identification of Germplasm Strategy) approach identifies traits related to drought adaptation in Vicia faba genetic resources. PLoS One 8(5): e63107.

Langmead, B., and S.L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9(4): 357.

Lipka, A.E., F. Tian, Q. Wang, J. Peiffer, M. Li, et al. 2012. GAPIT: genome association and prediction integrated tool. Bioinformatics 28(18): 2397–2399.

Lorenz, A., and K.P. Smith. 2015. Adding genetically distant individuals to training populations reduces genomic prediction accuracy in barley.

de Los Campos, G., J.M. Hickey, R. Pong-Wong, H.D. Daetwyler, and M.P.L. Calus. 2013. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193(2): 327–345.

de Los Campos, G., Y. Veturi, A.I. Vazquez, C. Lehermeier, and P. Pérez-Rodr\'\iguez. 2015. Incorporating genetic heterogeneity in whole-genome regressions using interactions. J Agric Biol Environ Stat 20(4): 467–490.

Mace, E.S., K.K. Buhariwalla, H.K. Buhariwalla, and J.H. Crouch. 2003. A high-throughput DNA extraction protocol for tropical molecular breeding programs. Plant Mol Biol Report 21(4): 459–460.

Maulana, F., K.-S. Kim, J.D. Anderson, M.E. Sorrells, T.J. Butler, et al. 2021. Genomic selection of forage agronomic traits in winter wheat. Crop Sci 61(1): 410–421.

McCormick, R.F., S.K. Truong, A. Sreedasyam, J. Jenkins, S. Shu, et al. 2018. The Sorghum bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. The Plant Journal 93(2): 338–354.

Meher, P.K., S. Rustgi, and A. Kumar. 2022. Performance of Bayesian and BLUP alphabets for genomic prediction: analysis, comparison and results. Heredity (Edinb): 1–12.

Meuwissen, T.H.E., B.J. Hayes, and M. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4): 1819–1829.

Moser, G., B. Tier, R.E. Crump, M.S. Khatkar, and H.W. Raadsma. 2009. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genetics Selection Evolution 41(1): 1–16.

Muleta, K.T., P. Bulli, Z. Zhang, X. Chen, and M. Pumphrey. 2017. Unlocking diversity in germplasm collections via genomic selection: a case study based on quantitative adult plant resistance to stripe rust in spring wheat. Plant Genome 10(3): plantgenome2016--12.

Muleta, K.T., T. Felderhoff, N. Winans, R. Walstead, J.R. Charles, et al. 2021. The recent evolutionary rescue of a staple crop depended on over half a century of global germplasm exchange. bioRxiv.

Nida, H., G. Girma, M. Mekonen, S. Lee, A. Seyoum, et al. 2019. Identification of sorghum grain mold resistance loci through genome wide association mapping. J Cereal Sci 85: 295–304.

Olatoye, M.O., L. V Clark, N.R. Labonte, H. Dong, M.S. Dwiyanti, et al. 2020. Training population optimization for genomic selection in miscanthus. G3: Genes, Genomes, Genetics 10(7): 2465–2476.

Olson, K.M., P.M. VanRaden, and M.E. Tooker. 2012. Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. J Dairy Sci 95(9): 5378–5383.

Passarella, V.S., R. Savin, and G.A. Slafer. 2008. Are temperature effects on weight and quality of barley grains modified by resource availability? Aust J Agric Res 59(6): 510–516.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M.A.R. Ferreira, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics 81(3): 559–575.

R Core Team. 2020. R: A Language and Environment for Statistical Computing.

Rakshit, S., K. Hariprasanna, S. Gomashe, K.N. Ganapathy, I.K. Das, et al. 2014. Changes in area, yield gains, and yield stability of sorghum in major sorghum-producing countries, 1970 to 2009. Crop Sci 54(4): 1571–1584.

Rhodes, D.H., L. Hoffmann, W.L. Rooney, T.J. Herald, S. Bean, et al. 2017. Genetic architecture of kernel composition in global sorghum germplasm. BMC Genomics 18(1): 15. doi: 10.1186/s12864-016-3403-x.

Sandhu, K.S., P.D. Mihalyov, M.J. Lewien, M.O. Pumphrey, and A.H. Carter. 2021. Genomic selection and genome-wide association studies for grain protein content stability in a nested association mapping population of wheat. Agronomy 11(12): 2528.

Sapkota, S., R. Boyles, E. Cooper, Z. Brenton, M. Myers, et al. 2020. Impact of sorghum racial structure and diversity on genomic prediction of grain yield components. Crop Sci 60(1): 132–148.

Shim, J., N.B. Bandillo, and R.B. Angeles-Shim. 2021. Finding Needles in a Haystack: Using Geo-References to Enhance the Selection and Utilization of Landraces in Breeding for Climate-Resilient Cultivars of Upland Cotton (Gossypium hirsutum L.). Plants 10(7): 1300.

Singh, R., and J.D. Axtell. 1973. High Lysine Mutant Gene (hl that Improves Protein Quality and Biological Value of Grain Sorghum 1. Crop Sci 13(5): 535–539. doi: 10.2135/cropsci1973.0011183X001300050012x.

Street, K., A. Bari, M. Mackay, A. Amri, and others. 2016. How the focused identification of germplasm strategy (FIGS) is used to mine plant genetic resources collections for adaptive traits. Enhancing crop genepool use: capturing wild relative and landrace diversity for crop improvement 54.

Stromberg, J.C., and J.A. Boudell. 2013. Floods, drought, and seed mass of riparian plant species. J Arid Environ 97: 99–107.

Tessema, G., H. Nida, A. Seyoum, M. Mekonen, A. Nega, et al. 2019. Ethiopian sorghum landrace SNP and phenotype data. doi: doi:/10.4231/PYQV-AT79.

Upadhyaya, H.D., and R. Ortiz. 2001. A mini core subset for capturing diversity and promoting utilization of chickpea genetic resources in crop improvement. Theoretical and Applied Genetics 102(8): 1292–1298.

Usai, M.G., M.E. Goddard, and B.J. Hayes. 2009. LASSO with cross-validation for genomic selection. Genet Res (Camb) 91(6): 427–436.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. J Dairy Sci 91(11): 4414–4423.

Velazco, J.G., D.R. Jordan, E.S. Mace, C.H. Hunt, M. Malosetti, et al. 2019. Genomic prediction of grain yield and drought-adaptation capacity in sorghum is enhanced by multi-trait analysis. Front Plant Sci 10: 997.

Wang, J., Z. Hu, H.D. Upadhyaya, and G.P. Morris. 2020. Genomic signatures of seed mass adaptation to global precipitation gradients in sorghum. Heredity (Edinb) 124(1): 108–121.

Whittaker, J.C., R. Thompson, and M.C. Denham. 2000. Marker-assisted selection using ridge regression. Genet Res (Camb) 75(2): 249–252.

Yu, X., X. Li, T. Guo, C. Zhu, Y. Wu, et al. 2016. Genomic prediction contributing to a promising global strategy to turbocharge gene banks. Nat Plants 2(10): 1–7.

Zhu, S., T. Guo, C. Yuan, J. Liu, J. Li, et al. 2021. Evaluation of Bayesian alphabet and GBLUP based on different marker density for genomic prediction in Alpine Merino sheep. G3 11(11): jkab206.

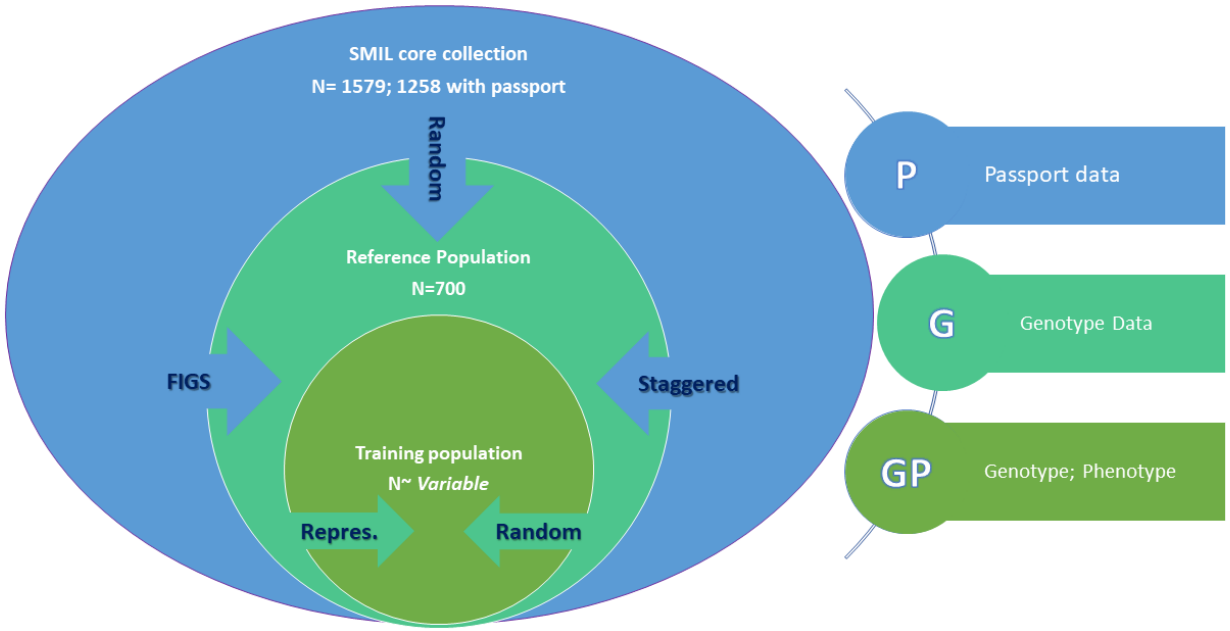
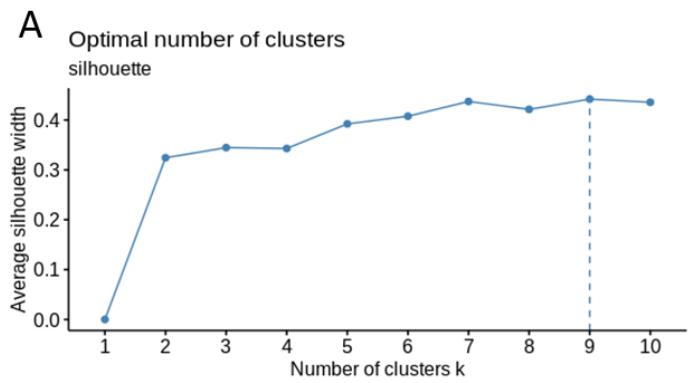


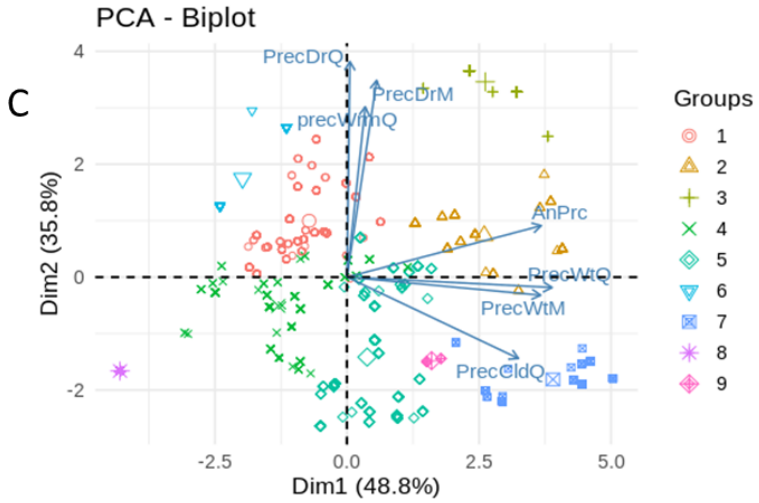
Figure 3-1 Schema showing how the reference and training populations were selected for this study.

The SMIL core collection is assumed to represent an extensive germplasm pool. Most of the genotypes are assumed with passport (n=1258) data. From these genotypes, reference populations (n=700) were sampled using Random, FIGS, or staggered approaches. The reference population is now assumed to be genotyped. Sampling of the reference population to form training population followed either random or representative (repres.) approach.



B

S/no	Cluster No	Average Ann Prec.	Size of Cluster	Cumulative
1	8	177	34	34
2	4	700	163	197
3	6	720	75	272
4	1	897	352	624
5	5	977	308	932
6	9	1140	5	937
7	2	1466	57	994
8	7	1519	70	1064
9	3	1650	63	1127



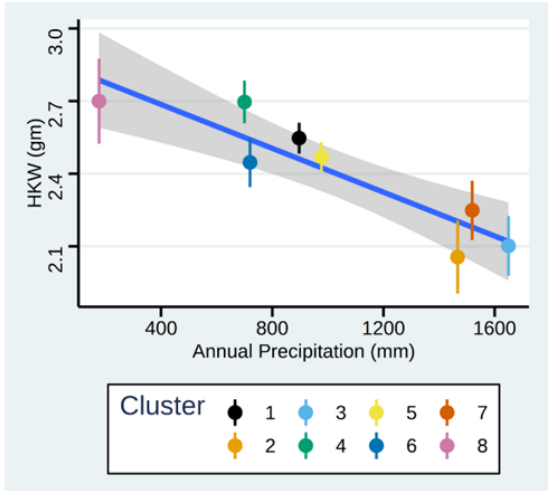


Figure 3-2 FIGS based selection of landraces: (A) Determination of optimum number of clusters (B) Group size and mean precipitation of the clusters. (C) The distribution of the clusters along precipitation gradient and PCs. (D) Relationship of mean cluster HKW and annual precipitation.

Abreviation: Ann. Annual, Prec. Precipitation, Dim- Principal component Axis, HKW- Hundred kernel weight

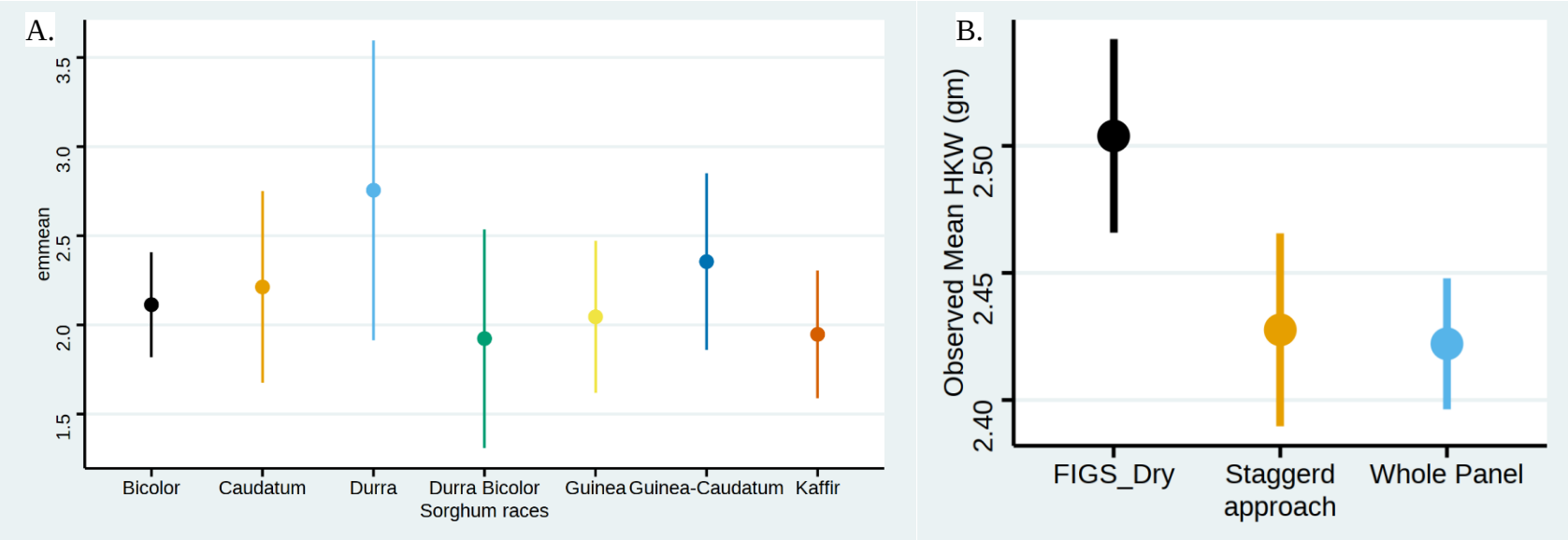


Figure 3-3 (A) Mean HKW performance of botanical races (bars show 95% confidence interval) (B) Mean HKW performance of reference populations drawn using different sampling strategies (bars shows 95% confidence interval).

Emmean: Adjusted mean of the group

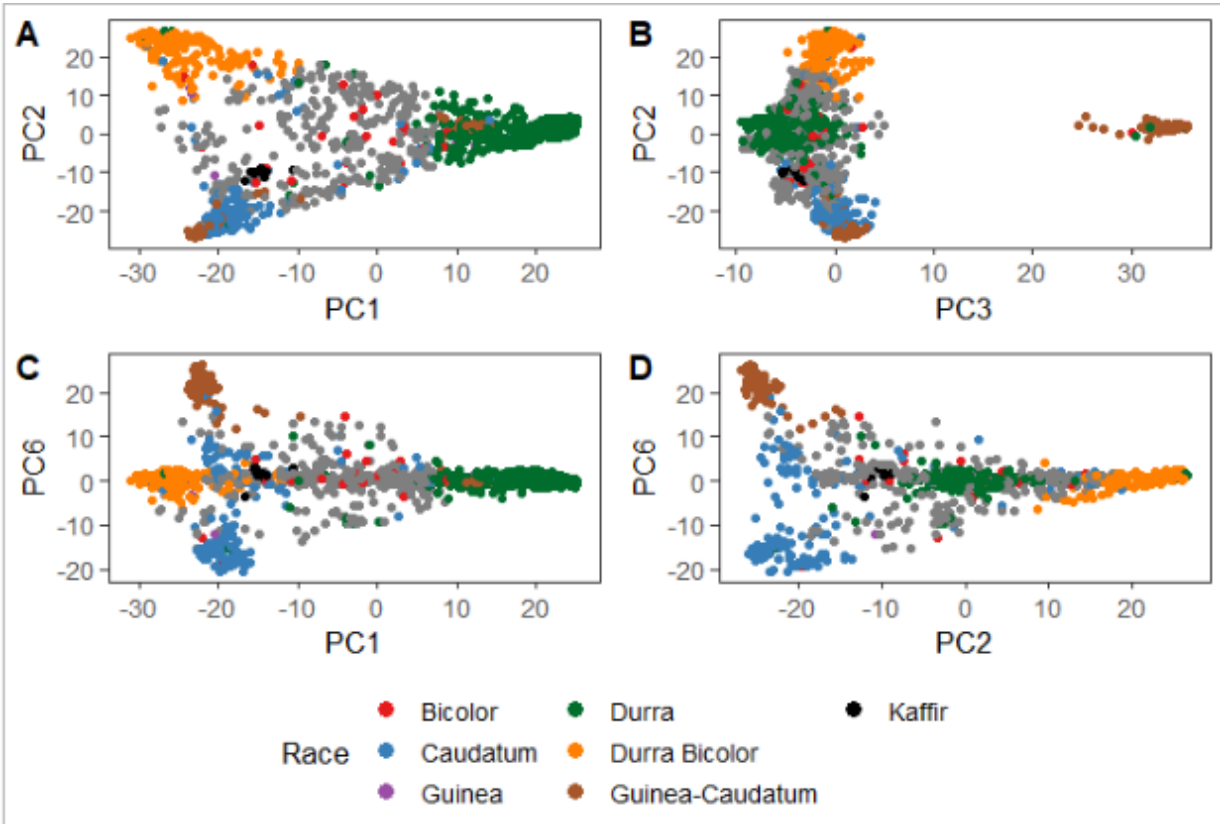


Figure 3-4 Population structure as evidenced by a few of the first PCs.

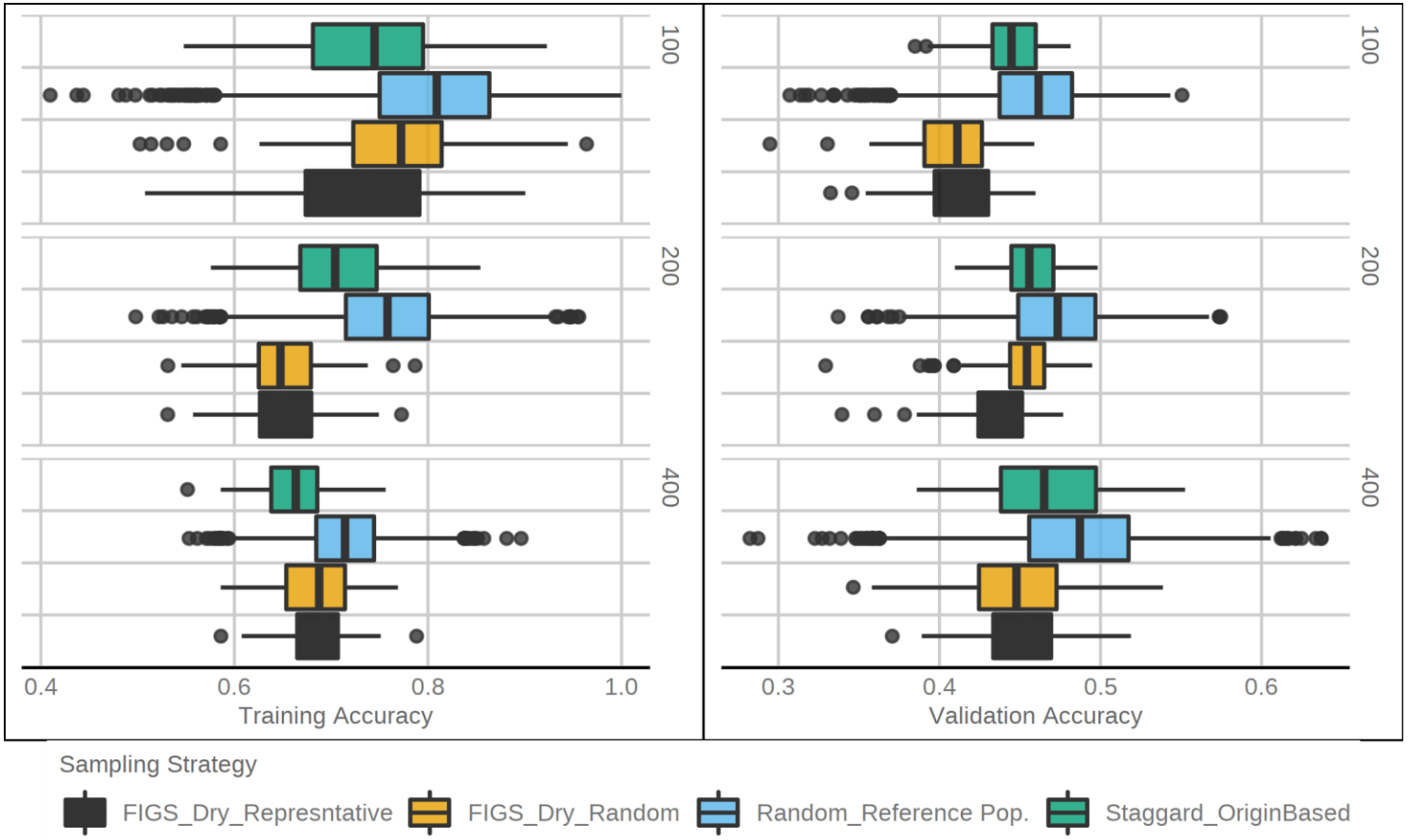


Figure 3-5 Training and validation accuracy of genomic prediction computed using reference populations assembled through FIGS (FIGS_Dry), Staggered (Staggered_OriginBased), and random (Random_Reference pop.) approaches.

FIGS_Dry_representative and FIGS_Dry_Random: FIGS approaches whose training individuals were selected using representative and random sampling approaches, respectively.

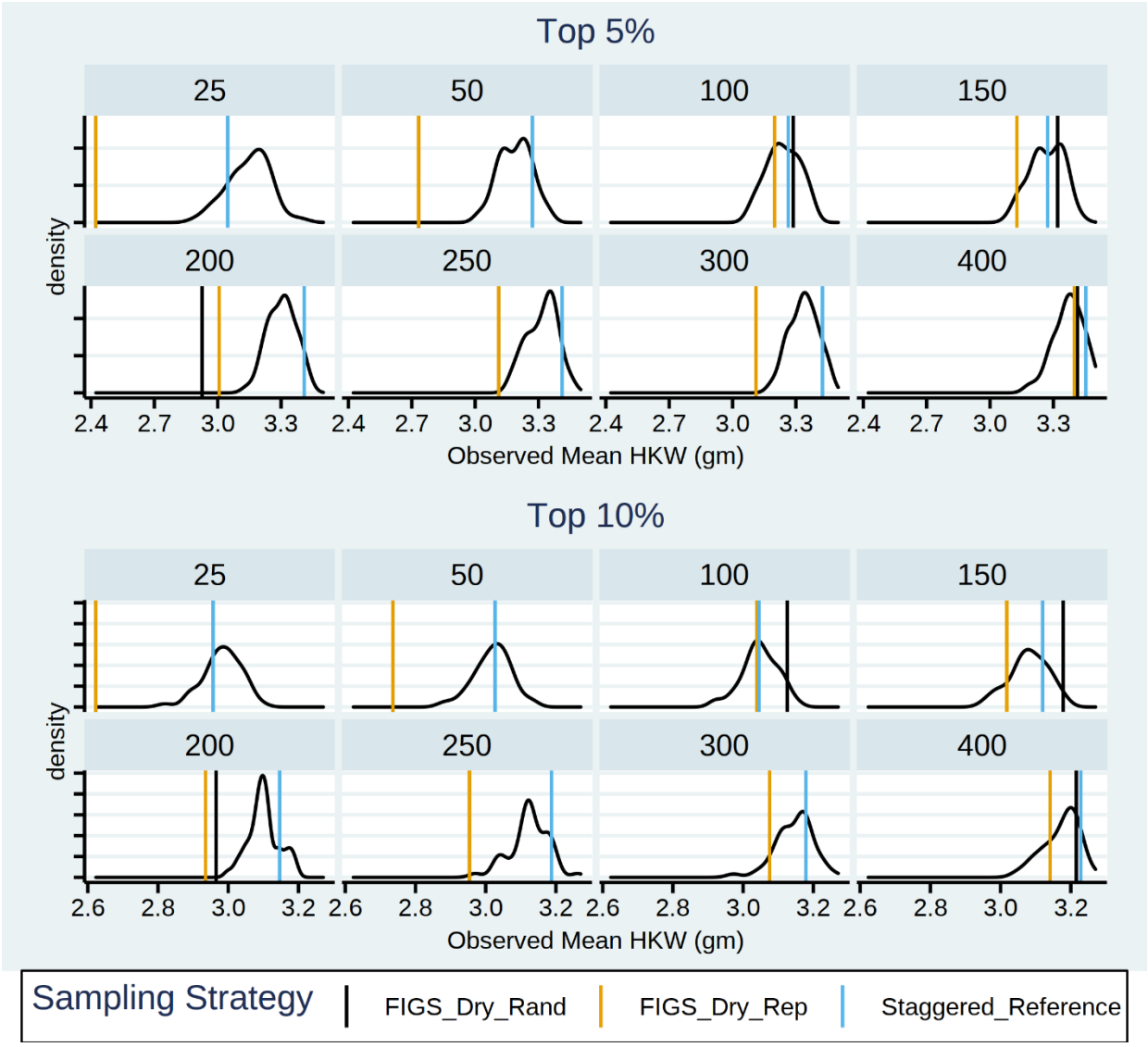


Figure 3-6 Mean HKW of top 5% and 10% lines selected based on GEBVs computed using different sampling strategies.

FIGS_Dry_rep and FIGS_Dry_Rand: FIGS approaches whose training individuals were selected using representative and random sampling approaches, respectively.

Table 3-1 Performance results of landraces evaluated in different environments.

Trait	Environment	Mean	Range	σ^2_g	σ^2_{res}	$H^{2\,(\%)}$
DTF	5	106.0	66.0 – 159.0	206.1	134.9	88.4
DTM	3	167.0	125.0 – 214.0	92.3	129.8	68.1
PH (cm)	6	317.0	97.9– 466.8	2972.7	1897.5	90.3
HKW (g)	2	2.4	0.5-4.6	0.2	0.2	66.1
YPP (g)	2	55.7	5.0 -174.0	287.2	540.2	51.5
GPC (%)	2	10.9	7.13 – 15.5	0.6	0.6	64.2

DTF: Days to Flowering; DTM: Days to Maturity; HKW: Hundred kernel weight; PH: Plant height; GPC: Grain protein content; YPP: Yield per plant.

Table 3-2 Prediction accuracy of landrace performance for six agronomic traits affected by training population size under rrBLUP using random samples drawn from the population.

	Prediction accuracy at different Training population sizes (n_{TP})								
Trait	n_{TP} =25	n_{TP} =50	n_{TP} =100	n_{TP} =150	n_{TP} =200	n_{TP} =250	n_{TP} =300	n_{TP} =400	n_{TP} =500
DTF	0.491	0.574	0.636	0.659	0.670	0.680	0.686	0.693	0.698
DTM	0.393	0.486	0.550	0.573	0.587	0.595	0.601	0.610	0.615
PH	0.486	0.551	0.597	0.619	0.630	0.638	0.645	0.654	0.661
HKW	0.395	0.426	0.448	0.460	0.468	0.473	0.477	0.483	0.489
GPC	0.225	0.272	0.298	0.310	0.317	0.323	0.326	0.331	0.333
YPP	0.212	0.273	0.316	0.341	0.354	0.364	0.371	0.381	0.389

DTF: Days to Flowering; DTM: Days to Maturity; HKW: Hundred kernel weight; PH: Plant height; GPC: Grain protein content; YPP: Yield per plant.

Table 3-3 Prediction accuracy of landrace performance for six agronomic traits affected by training population size under GBLUP using random samples drawn from the population.

Prediction accuracy at different Training population sizes (n_{TP})								
Trait	n_{TP} =25	n_{TP} =50	n_{TP} =100	n_{TP} =150	n_{TP} =225	n_{TP} =300	n_{TP} =400	n_{TP} =500
DTF	0.506	0.565	0.662	0.672	0.675	0.690	0.697	0.699
DTM	0.418	0.520	0.531	0.570	0.602	0.599	0.611	0.618
PH	0.476	0.518	0.608	0.613	0.640	0.646	0.653	0.665
HKW	0.406	0.425	0.455	0.459	0.472	0.477	0.473	0.488
GPC	0.209	0.245	0.305	0.301	0.319	0.322	0.329	0.333
YPP	0.220	0.260	0.317	0.335	0.365	0.373	0.374	0.396

DTF: Days to Flowering; DTM: Days to Maturity; HKW: Hundred kernel weight; PH: Plant height; GPC: Grain protein content; YPP: Yield per plant.

Table 3-4 Composition of reference populations established based on origin information.

s/no	Race	Total	Expected Proportion1	FIGS Selected	FIGS Selected Proportion	FIGS Population Change from base	Staggered Selected	Staggered Selected Proportion	Staggered Population Change form base
1	Bicolor	35	0.028	13	2.3%	-14.9%	18	3.1%	13.88%
2	Caudatum	195	0.153	50	9.0%	-41.3% ***	61	10.6%	-30.73%**
3	Durra	617	0.485	309	55.7%	14.7% ***	292	50.9%	4.79%
4	Durra-bicolor	237	0.186	105	18.9%	1.5%	131	22.8%	22.39% *
5	Guinea	6	0.005	1	0.2%	-61.8%	3	0.5%	10.71%
6	Guinea-Caudatum	171	0.135	76	13.7%	1.8%	64	11.1%	-17.13%
7	Kaffir	10	0.008	1	0.2%	-77.1%	5	0.9%	10.71%

(*), (**), (***) show statistically significant different proportions from the expected proportion evaluated using the binomial test for equality of proportions.

Table 3-5 Characteristics of reference populations sampled using different approaches relative to the whole panel used.

Parameter	FIGS	Staggered	Whole Panel
LD decay to half maximum	~3 kbp	~3 kbp	~2.5 kbp
LD decay to r^2 0.1	~59 kbp	~63 kbp	~40kbp
Average pairwise Distance (on basis of IBS)	0.303	0.306	0.309

Table 3-6 Geographic descriptions of study sites and number of data points collected for each trait from each location (data include un-genotyped individuals).

No	Site	Location	Mean Temperature* (^{0}C)	Mean precipitation* (mm)	The Koppen-Geiger climate classification*	Number of data points used*					
						PHT (Year)	DTF	DTM	HKW	GPC	YPP
1	Arsi Negele	7°21' N, 38°42' E	18	915	humid subtropical climates	1943	1629		1761 (2016)	781	1326 (2016)
2	Haramaya University	9°24' N, 42°01' E	18	799	Subtropical highland climate	1983; 1520 (2016)	1944	1972	1498 (2016)	405	1215 (2016)
3	Meiso	8^{0}59'N, 40^{0}25'E	23.	831	Tropical Climate	1301	1992	1964			
4	Pawe	11° 18'N, 36° 24E'	24	1601	Tropical Climate	1960	2008				
5	Bako	9° 05'N, 37° 02E'	20	1200	Tropical wet and dry or savanna climate	1998	1998	1998			

*Reference:(Fick and Hijmans, 2017; Tessema et al., 2019). The number of data points without an accompanying bracket represents 2015 data

Chapter 4 - Genotype and pre-processing treatments impact in-vitro protein digestibility (IVPD) in the Ethiopian fermented bread from sorghum

Abstract

Sorghum is an important source of calorie and protein for smallholder farmers in Sub-Saharan Africa and Southeast Asia. Nevertheless, the low digestibility of sorghum proteins and the lack of access to alternative protein sources make consumers vulnerable to protein malnutrition. Processing sorghum into cooked products generally reduces protein digestibility with fermented products tending to suffer less compared to unfermented products. The objective of this study was to investigate the impact of grain processing treatments on the protein digestibility of the Ethiopian fermented flatbread. Four sorghum genotypes (TxArg-1, B503, Macia, and Dorado) were subjected to four processing treatments, decortication, sprouting, roasting, and unprocessed control, and milled to two-particle sizes resulting in a total of 32 treatments. The genotype, processing treatment, and their interaction were significant for in vitro protein digestibility (IVPD) in cooked and uncooked samples. The roasting treatment significantly reduced IVPD compared to the unprocessed control while sprouting significantly increased IVPD. Decortication appears to have no impact on IVPD. Finer particle size tends to enhance IVPD in all genotypes and for all processing methods. Processing treatments slightly affected protein content, with the trait significantly improved by the sprouting treatment.

Introduction

Sorghum (*Sorghum bicolor (*L.) Moench) is one of the five major cereal crops of the world. The developed nations utilize sorghum primarily as animal feed (Ronda et al., 2019) while the crop is cultivated as principal food crop in developing countries of Sub-Saharan Africa and South Asia (Ratnavathi and Patil, 2013). Inherent characteristics of the crop such as resilience to drought, low input requirements, and long tradition of its use and cultivation, makes sorghum widely preferred by smallholder communities in developing countries. Although sorghum is similar to other cereals in terms of nutritional composition, the availability of nutrients, especially proteins in sorghum-based diets, appears to be low, making the consumers vulnerable to protein malnutrition (Maclean et al., 1981; Semba, 2016).

The digestibility of sorghum proteins is a complex process involving several factors, with the major ones related to the characteristics of the sorghum protein itself (Duodu et al., 2003). Sorghum storage proteins, the kafirins, are organized into spherical protein bodies with enzyme-resistant fractions often occurring on the outer layer restricting access of digestive enzymes to the more digestible kafirin fractions (Duodu et al., 2003). Moreover, starch granules are embedded into the protein matrix, rendering sorghum starch recalcitrant to enzymatic digestion (Rooney and Pflugfelder, 1986). In cooked sorghum, gelatinized starch may also inhibit the digestion of sorghum proteins (Duodu et al., 2002). Several factors in sorghum extrinsic to storage proteins but inherent to the grain characteristics may also influence in-vitro protein digestibility, IVPD (Duodu et al., 2003). Sorghum grain carries trypsin proteinase inhibitors in its bran layer, which may lower protein digestibility (Kumar et al., 1979). Sorghum also contains phytate, which can form protein-phytate complexes (Kumar et al., 2010). Certain types of sorghum also have tannins that bind with proteins and reduce their digestibility.

Efforts to improve the nutritional quality of sorghum in the past have focused on addressing these factors and have made significant progress towards understanding the structure, chemistry, and genetic factors behind sorghum protein digestibility and reducing the adverse effects of anti-nutritional factors. Germplasm lines with improved protein digestibility have been developed, though their commercial deployment is pending (Weaver et al., 1998; Tesso et al., 2006). However, there are additional factors that need to be addressed to remove the barrier that low

protein digestibility imposes on the value of the crop, both as animal feed and human food. Food processing methods and genotypes have been shown to affect the digestibility of proteins from sorghum foods (Weerasooriya et al., 2018; Rashwan et al., 2021). Cooking sorghum food products, primarily through wet process, reduces protein digestibility dramatically, with the degree of reduction varying by food types and processing methods (Weerasooriya et al., 2018). Furthermore, there is little or no correlation between protein digestibility from raw flour and cooked foods which complicates improving the trait for human food since it requires samples to be cooked to screen them for digestibility (Weerasooriya et al., 2018).

Smallholder farmers in Sub-Saharan Africa heavily depend on sorghum as the primary source of protein and energy (Belton and Taylor, 2004). According to FAOSTA (2018), the population in the major sorghum-consuming country of Sudan derived 24% of its dietary protein from sorghum (Chavan et al., 1988). Disaggregation of consumption data by occupation or economic opportunity would undoubtedly show the percentage to be higher among poor rural communities (Gali and Rao, 2012).

Sorghum is traditionally consumed after processing into different food types. The most common sorghum-based foods used in the developing world include porridge, and fermented and unfermented breads, all of which involve wet cooking and cause a substantial reduction in protein digestibility. Given that smallholder farmers do not have access to animal protein to supplement their diet, the low availability of protein from this staple cereal exposes the community to severe protein malnutrition. The prevalence of stunting and wasting reported widely among rural children in sub-Saharan Africa is evidence of severe protein-energy deficiency (Semba, 2016; Derso et al., 2017; Gerald J. and Dorothy R. Friedman, 2019; Gebre et al., 2019; Gebreegziabher and Regassa, 2019).

Fermented flatbread, recognized by different names in different communities, is the major component of local dishes in Ethiopia and Sudan. While the best quality of such bread in Ethiopia is made from teff, sorghum comes second and exhibits the same quality as teff bread when mixed with teff up to 50%. Communities in sorghum-producing regions in Ethiopia prefer sorghum bread to teff. Because teff has primarily become a cash crop, rural communities in teff-growing areas such as the central highlands use teff-sorghum composite flour. The increasing price for teff grain

in recent years, partly forced by the growing foreign market and population pressure, is forcing a large segment of the urban community to turn to sorghum as alternative food grain making it the third most widely used food crop in Ethiopia (Taffesse et al., 2013; Demeke and Di Marcantonio, 2019). Given the changes to teff utilization and sorghum consumption, the need to improve protein digestibility in sorghum becomes even more important to protect communities from hidden protein malnutrition.

The traditional bread-making process in Ethiopia involves fermentation of the dough. Fermentation enhances protein availability by prompting structural changes in sorghum storage proteins, kafirins (Taylor and Taylor, 2002). Previous studies have shown that fermentation improves protein digestibility (Correia et al., 2010), alters starch properties (Abd Elmoneim et al., 2017), and enhances the overall nutritional quality of food products (Chavan et al., 1988; Osman, 2004; Weerasooriya et al., 2018; Adeyeye et al., 2019). A comparison of fermented and unfermented sorghum food supplements in laboratory rats showed that fermented sorghum supplements improved nutritional parameters (Adejuwon et al., 2020). Although cooking generally reduces protein digestibility, the extent of the reduction is lower in the fermented product, and uncooked fermented sorghum has higher IVPD relative to unprocessed raw sorghum (Weerasooriya et al., 2018). Even though fermentation enhances the IVPD of uncooked samples, digestibility was still compromised by wet cooking relative to uncooked grain, which is an integral part of bread making in many traditional food products, including the Ethiopian fermented bread *injera* (Taylor and Taylor, 2002) and in many food products made from sorghum. This suggests that opportunity still exists to enhance the digestibility of proteins from cooked sorghum foods using pretreatment procedures. The objective of this study was to optimize the bread-making processes and methods to improve protein digestibility in the Ethiopian fermented flatbread.

Material and methods

Plant materials

Four sorghum genotypes, B503, TxArg1, Macia, and Dorado, were used in this study. B503 is a recent seed parent line release from Kansas State University. The grain has typical burgundy color and is high in total protein and lysine. TxArg-1 is an old white seeded waxy sorghum used

in various food quality studies. Macia and Dorado are both white seeded food-grade sorghums widely grown and used in various types of foods in Africa. Seeds of these genotypes were increased during the 2017 growing season at Kansas State University Agronomy Research Farm, Ashland Bottoms, near Manhattan, Kansas. At maturity, the grains were harvested, cleaned and stored until use.

Experimental design and treatments

The experiment consisted of three factors, genotype, and processing treatments each consisting of four levels, and flour particle size consisting of two levels. The genotype factor consisted of four diverse sorghum genotypes (B503, TxArg1, Macia, Dorado) described above. The four levels of processing factor included sprouting, decortication, roasting and the unprocessed control. The two levels of flour particle sizes were achieved by grinding the samples using 0.5 mm and 2 mm screen sizes. The treatment was a factorial combination of the different levels of these factors totaling to 32 treatments. Grain samples of genotypes from the same batch were sampled and divided into four subgroups for processing treatments (sprouting, decortication, roasting and control). The experiment was conducted in randomized complete block design with two replications, where replications were considered as blocks. Due to large treatment size, the experiment was carried out at two different times representing the two replications.

Grain processing procedure

Sprouting

Grains were rinsed with tap water and placed in a 10 cm deep 20 cm diameter metallic pan. Approximately 2.5 liter of tap water was added to the pan to fully submerge the grains with the water level visibly above the grains and was maintained this way for 48 h with the water replaced every 8 to 10 h. The samples were then transferred to 35 cm × 21cm × 12 cm trays with multiple layers of a water-soaked paper towel placed at the bottom of the tray. Additional paper towels were used to fully cover the grains. The trays were left in the dark at room temperature for another 48 h. The grain layers were intermittently mixed to insure uniform germination. Moldy-looking grains

were removed as soon as observed. After 48 h of germination, the samples were spread thin on flat metallic pan to air dry.

Decortication

The decortication treatment was conducted using a facility at the Kansas State University Grain Science Department. A custom-made sample abrasive decorticator STRONG-SCOTT-17810 (Minneapolis, MN) fitted with a 6-inch diameter and 24 grit grinding disc on a 0.25 hp motor was used. Five-hundred grams of grain of each of the genotype was subjected to decortication. The machine was set to remove twenty percent of the bran layer. Due to the inherent differences in grain characteristics, the average decortication time for each genotype was different with TxArg1, B503, Dorado, and Macia requiring 2:30, 2:45, 2:50, and 2:45 mins, respectively.

Roasting

The roasting treatment was applied using a clay pan preheated to 100 °C. The grains were poured onto a hot pan and roasted for five minutes with occasional stirring to achieve uniform exposure to the surface of the pan. The samples were then allowed to cool down on a kitchen counter for one hour.

Control

These are intact grain samples of each genotype not subjected to any processing treatment. All processed samples were placed in a labeled plastic container for storage at room temperature.

Sample Grinding

The sprouted, roasted, decorticated, and intact control samples were ground into the two selected particle sizes using a UDY Cyclone Sample Mill (Udy Corporation, Fort Collins CO). Each of the pretreated samples was divided in half, with one half ground using a 0.5 mm screen and the other half with 2 mm screen providing two different particle sizes. The flour samples were collected into a labeled Ziploc bags and stored at 4°C until needed. The particle size distribution of the flour was later analyzed in duplicates using LS 13 320 Particle Analyzer (Beckman Coulter

Life Sciences, Indiana, USA) with Universal Liquid Module. Particle sizes were reported as volume-weighted mean diameter.

Food sample preparation

The sample foods (fermented bread) were prepared from each of the thirty-two treatment combinations using the following procedure. About 200 g flour from each sample were mixed with 500 ml water along with a 10 ml of liquid commercial yeast (5%). The mixture was left at room temperature to ferment for 48 h. Then the samples were placed in the refrigerator to stop further fermentation. About 30 g of the fermented dough was taken in a metal cup, mixed with 100 ml water, and heated on a stove until gelatinized. The gelatinized dough was added back to the original dough and thoroughly mixed. Warm water was added until the dough was soft enough for spreading on a pan and was let to set for two hours before cooking into bread. For cooking, a 40 cm wide, circular electrical claypan (WASS Electronics Inc., VA, USA) was preheated to 121°C. The slurry was spread circularly on the pan and covered with a metal lid to cook for two minutes. The cooked samples were then cooled down, lyophilized, and ground for analysis.

Sample characterization

IVPD was determined in duplicates from each of the pretreated raw flour and fermented food samples using the method of Mertz et al. (1984) as detailed in Cremer et al. (2014). Protein content (PC) of unprocessed raw samples, and processed samples before and after cooking measured by nitrogen combustion method using a TruSpec CN combustion analyzer (LECO Corp, St. Joseph, MI). The nitrogen content was multiplied by the 6.25 factor to obtain crude protein content estimate. Change in protein content (ΔPC) of processed and cooked samples with reference to the unprocessed raw samples.

Statistical analysis

Model fitting was conducted using the lm function of the R software (R Core Team, 2020). Analysis of variance (ANOVA) was performed using the Anova function from the car package (Fox and Weisberg, 2018). The functions emmeans from the package emmeans (v 1.5.2, (Lenth et

al., 2019)) and cld from the multcomp package (v 1.4-17 (Hothorn et al., 2016)) were used for marginal means comparison. Type I error (α) level was set at 0.05 with Bonferroni adjustment for multiple testing. Paired t-test was conducted using t-test function from the R package (v 4.0.3, (R Core Team, 2020).

Results

The genotypes used in this study possess diverse physicochemical attributes. Genotype B503 had the highest PC of 14.4 % and the highest proportion of vitreous endosperm (Figure 4-1 and Table 4-1). The African cultivar Macia had the lowest PC of 12% and an intermediate vitreosity comparable to Dorado, another African cultivar. TxArg-1 was a waxy genotype with near zero percent apparent amylose content while others are non-waxy with amylose content of about 14% (Table 4-1). All genotypes except B503 have white pericarp color. Figure 4-1 B shows the flour particle size distribution of the genotypes after milled with a 2mm screen. Particle sizes ranged from near zero to 1600 μm and peaked around 500 μm for B503 and Dorado, while BTxArg1 and Macia had lower particle sizes as their peak (Figure 4-1 B).

Analysis of variance

The analysis of variance for the effects of the different factors on IVPD and PC in cooked and uncooked states is presented in. Both traits were highly significant for all factors, except the genotype effect for raw IVPD and raw PC for particle size (P.value < 0.01). The interaction between these parameters and the three-way interaction was not significant for both raw PC and raw IVPD. Likewise, all three factors were highly significant for both IVPD and PC in the cooked samples (P.value < 0.01). The interaction between genotype and processing methods was also highly significant for both IVPD and PC. However, the processing method by particle size interaction was significant only for PC. IVPD and PC for Genotype by particle size interaction for both IVPD and PC, processing method by particle size for IVPD and the three-way interaction both for IVPD and PC were not significant (Table 4-2).

IVPD and PC in uncooked samples

The pre-processing treatments were shown to significantly affect IVPD and PC in uncooked samples of all test genotypes. Decortication significantly reduced PC while other processing treatments, sprouting, and roasting did not have substantial effect on PC (Table 4-3). Likewise, PC was significantly different among the test genotypes with B503 having the highest uncooked PC (14.3%) while genotype Macia had the lowest (11.9%). The effect of particle size on PC, however, was not significant (Table 4-2).

The effect of pre-processing treatment on IVPD of uncooked samples was also significant (Table 4-2 and Table 4-4). Unlike the PC, where decortication was shown to have the most considerable effect, roasting was shown to have the highest negative effect on IVPD in all genotypes (Table 4-4). On the other hand, sprouting significantly improved IVPD in all genotypes. The decortication process did not have significant effect on IVPD unlike the PC. Compared to the control treatment, the reduction in IVPD across genotype and particle size in roasted samples was 44.2%, while the improvement in IVPD in sprouted samples compared to the untreated control was 10.2%. Unlike the PC, IVPD was significantly higher in finer flour particles than the coarser flours in all genotypes. IVPD in uncooked samples was not significantly different between genotypes.

IVPD and PC in cooked samples

Similar to the uncooked samples, IVPD and PC in cooked samples were significantly affected by processing treatments, genotypes, and flour particles. Roasting significantly reduced PC while sprouting improved PC compared to the unprocessed control (Table 4-5). Overall, sprouting improved PC by 11% and there is significant difference between genotypes. Relative to the control, sprouting increased PC by 11.8% in TxArg1, 8.7% in Macia, and by 6.1% and 16%, in B503 and Dorado, respectively. On the other hand, roasting reduced PC by 6% while decortication did not have significant effect.

Across processing treatments, B503 had the highest PC of 14.7% in the cooked state, followed by TxArg1 with 14.3% PC. The lowest mean PC was reported in the African genotype Macia while the other African genotype Dorado was about average between the high and low

genotypes. Flour particle size also significantly affected PC of cooked samples in all genotypes and processing treatments. Unlike in the uncooked samples, food samples prepared from smaller particle size flours consistently had higher PC than those made from coarser flour (Table 4-5).

As expected, IVPD in cooked products was markedly lower than in the uncooked products. B503 had the lowest reduction of 29% while Macia had the highest reduction (45%). Dorado and TxArg-1 had 35% and 38% reduction in IVPD compared to the uncooked samples. IVPD in the cooked products was also significantly affected by processing treatments, genotypes, and interaction between them (Table 4-2 and Table 4-6). Roasting treatment caused significant reduction (46%) in IVPD compared to the unprocessed control, while sprouting significantly improved (40%) IVPD. The reduction in IVPD in roasted samples relative to their respective control was 22.6, 54.9, 52.5, and 53.2% in B503, Dorado, Macia, and TxArg-1, respectively, while sprouting accounted for 47.4, 53.5, -4.7, and 65.6% increase in B503, Dorado, Macia, and TxArg-1, respectively. Decortication, although assumed to have removed the bran layer that is believed to carry various anti-nutritional factors, did not significantly affect IVPD.

Among genotypes, the highest marginal mean of cooked IVPD of 32.6% was obtained in B503, followed by the waxy genotype TxArg-1 (31.6%), while the least (25.1%) was found in the African cultivar, Macia. IVPD of cooked samples was also affected by the interaction between genotypes and processing treatments, indicating that the IVPD response of genotypes to processing treatments was different (Table 4-2 and Table 4-6). Although roasting reduced IVPD across genotypes, the response of individual genotypes to roasting was different and was similar for sprouting, contributing to the significant interaction effect of the genotype × processing method. Although not of similar magnitude to the uncooked samples, flour particle size had significant effect on IVPD in cooked sampled with food samples from finer particles are on average 8.8% higher in IVPD than those from coarser flowers (Table 4-4 and Table 4-6).

Changes in PC associated with processing treatments

Cooking sorghum has been reported to undermine protein availability to a variable degree depending on cooking methods and processing. The processing of sorghum into different food products has been reported to affect the availability of nutrients, especially of proteins. In this

study, we compared the effect of cooking on sorghum samples subjected to different processing treatments. In general, cooking doesn't seem to have a negative impact on PC but food samples from sprouted grain had significantly higher ΔPC in all genotypes while roasting tended to reduce PC in some genotypes (Figure 4-2).

The other area to draw a comparison is the effect of processing treatments on IVPD. Regardless of the processing treatments subjected to, cooked products have markedly less IVPD than their respective uncooked samples. Across genotype mean IVPD was reduced by 41.3% in cooked samples subjected to decortication treatment while the roasting and sprouting treatments had 43.3 and 23.6% reductions in the IVPD of their cooked samples. The across genotype reduction in IVPD in the control samples was 29.9%. In other words, IVPD in the cooked control treatment sample was only 70.1% as high as in the uncooked sample. The IVPD from the uncooked control treatment was 47.4, 48.2, 52.4, and 54.2% for B503, Dorado, Macia, and TxArg1, respectively. However, when cooked into fermented flatbread, the IVPD dropped to 29.7, 28.6, 31.6, and 29.9, respectively, showing that protein in the cooked samples were only 62.7, 59.3, 60.3, and 55.2% as digestible as those in the uncooked samples (Table 4-7). Hence the goal of the research was to identify a processing treatment that can increase this proportion. The data in this study revealed that the processing treatments of decortication and roasting did not provide any benefit but the sprouting treatment significantly improved IVPD of cooked products. Compared to the untreated control, the IVPD of B503, Dorado, and TxArg-1 improved to 92.4, 91.1, and 91.3% in cooked food samples prepared from sprouted grains compared to 62.7, 59.3 55.2% in cooked foods from untreated samples. IVPD in genotype Macia, however, did not show positive response to sprouting.

Discussion

Because sorghum is uniquely low in the bioavailability of its proteins, especially when cooked, protein digestibility has been the focus of numerous studies on this crop (Duodu et al., 2003). Although different processing methods had been shown to improve cooked protein digestibility, all of them were low compared to digestibility values from raw grain samples. In a previous study, our group investigated IVPD in various food products commonly consumed in Africa, where the fermented flatbread from Ethiopia was shown to have markedly higher IVPD

compared to other unfermented food products (Weerasooriya et al., 2018). Thus, the focus of this study was to optimize the processing method for making the Ethiopian flatbread that can enhance the digestibility of sorghum proteins without significantly altering the taste and texture of the product. Based on previous research, we expected that genotypes might react differently to processing treatments. The results may help fine-tune breeding goals towards setting tailored breeding objectives to select cultivars more suitable for making fermented bread with higher protein digestibility. In this study, we produced test samples from four tannin-free sorghum of African and temperate origin varying in seed color, starch properties, and protein profile and three-grain pretreatment procedures along with untreated check milled at two particle sizes. The samples were then tested for protein digestibility both before and after cooked to the Ethiopian fermented bread.

Processing treatment significantly affected the PC of uncooked samples (Table 4-2). Decortication that involved removal of the bran layer reduced mean PC significantly compared to the unprocessed control (Table 4-2). It appears that in addition to removing the pericarp layer, decortication may also partly remove the protein-dense germ and peripheral endosperm layers leading to an overall decline in PC of samples (Yetneberk et al., 2005). However, both roasting and sprouting did not have a significant effect on PC of uncooked samples (Table 4-3). However, Previous studies on other crops also showed sprouting increasing PC (Warle et al., 2015). Among the likely factors attributable to this is the respiratory loss of carbohydrates through CO_2, effectively concentrating protein in the remaining grain (Dicko et al., 2006). Several other studies reported results countering this finding (Subramanian et al., 1995; Elkhalifa et al., 2010; Afify et al., 2012; Singh et al., 2017; Yi et al., 2017). However, in many of these studies, the sprouts were removed before evaluating grain PC, which may have resulted in the loss of a significant portion of protein and, therefore, a decline in PC (Subramanian et al., 1995; Afify et al., 2012). Taylor (1983) also reported that germination induced the partitioning of a significant portion of protein and non-protein nitrogen to the developing roots and shoots. In the current study, the sprouts were dried and milled together with the grain and thus there was minimal loss of nitrogen in the process.

PC in the cooked bread was higher than the respective uncooked sample PC. A paired t-test between pairs of raw and fermented and then baked bread showed that baked samples had significantly higher PC than raw samples ($P < 0.01$). Such an increase in PC in fermented products

is consistent with the increase in PC in yeast-mediated fermented products reported by Day and Morawicki (2016). The fermentation process in the traditional preparation of Ethiopian bread is mediated by a complex mixture of yeast and lactic acid bacteria (Tadesse et al., 2019). Yeast-mediated fermentation is responsible for the characteristic honey-comb-like holes in the bread, and the release of CO_2 from the slurry's air pockets is responsible for carving out of the "eyes" (Attuquayefio, 2014) and, as a result, concentrated protein through depletion of carbon from the system (Day and Morawicki, 2016).

This study determined ΔPC, the difference between the PC of cooked bread and the PC of respective uncooked samples. In our study, sprouted samples had the highest (ΔPC) (Figure 4-2B), and this may be due to enhanced fermentation in sprouted samples and the resulting disproportionate mass loss in the form of CO_2. Sprouting had been shown to increase reduced sugar (Mella, 2011) and free amino nitrogen (Yi et al., 2017) concentrations in samples, both of which are rate limiting inputs of fermentation, and their abundance in sprouted samples may have fueled the fermentation step in the bread making process (Pickerell, 1986).

Fine milling also appeared to increase PC and ΔPC in cooked samples (Table 4-5 and Figure 4-2 B). Generally, samples ground using the finer 0.5 mm screen had higher PC (Table 4-5). Similarly, except for the roasted sample, where ΔPC was essentially zero, ΔPC was higher for finely milled sprouted samples than for courser samples (Figure 4-2 A). This may be due to the likely increase in the fermentation rate and the resulting increase in PC as aforementioned. Fine milling increases the surface area for enzymatic action (Mahasukhonthachat et al., 2010). Higher reducing sugar concentration was achieved in finely milled sorghum through enhanced amylase-mediated starch digestion (Barcelos et al., 2011). The improved accessibility of starch and other supplies have enhanced the overall fermentation rate and as a result increased PC and ΔPC. The zero ΔPC in roasted samples in both milling sizes may indirectly show that fermentation rate in the roasted samples might have been the slowest.

Processing treatments, genotypes, and particle size significantly affected protein digestibility both in cooked and uncooked states; however, not all processing treatments and genotypes had a significant effect on the trait (Table 4-2). The effect of decortication on bread IVPD relative to the control was not significant (Table 4-6). Although removing the bran layer

through decortication was expected to remove anti-nutritional compounds including phenols and protease inhibitors (Nikmaram et al., 2017), this did not significantly affect IVPD in the current study. Previous studies have shown that non-tannin phenolic compounds such as flavonoids and phenolic acids did not significantly impact protein digestibility (Duodu et al., 2003; Emmambux and Taylor, 2003). However, these compounds are potent inhibitors of the activity of alpha-amylase (Funke and Melzig, 2005). Another study on non-tannin sorghums showed that decorticating the grain increased the reducing sugar production and overall fermentation rates (Alvarez et al., 2010). Contrary to the expectations, decortication reduced IVPD in cultivar Macia (Table 4-6). This is difficult to explain but it may have to do with the elimination of the more digestible protein from the germ (Alvarez et al., 2010). Also, Macia is white, tan plant containing likely low amount of anti-nutritional factors in the bran that removal of the brain layer has little or no impact on IVPD.

The other processing treatment, roasting, rather had a significant effect on IVPD, with all genotypes showing a significant reduction in IVPD both before and after cooking. It is not clear what chemical changes have occurred due to roasting that increased resistance to pepsin digestion. In another report by our group, roasted food products were shown to have higher IVPD compared to fermented products (Weerasooriya et al., 2018). However, the roasted products in the previous study involved a prolonged heat treatment in achieving dry cooking. In contrast, the roasting in the current study was imposed to simulate the traditional practices in Africa, where the grain is subjected to light roasting aimed at facilitating drying and milling instead of cooking. In other words, the degree of roasting used in the current study, unlike the previous one, was only partial and did not achieve complete cooking. Nevertheless, the reduction in IVPD following partial roasting needs further investigation.

Unlike decortication and roasting, sprouting significantly increased protein digestibility in most of the genotypes, especially in the cooked state. In B503, Dorado, and TxArg-1, sprouting increased bread IVPD by 47.5%, 53.5 %, and 65.6 %, respectively, relative to the unprocessed cooked control sample. It is not clear what factor (s) contributed to improved IVPD in sprouted samples. It appears that sprouting activated starch hydrolyzing enzymes softened the endosperm making it more prone to fermentation, which is part of the cooking process. Fermentation appears to further degrade the starch granules exposing the protein bodies to action by protease enzymes.

Previous studies have reported on the synergistic effect of sprouting and fermentation on improving IVPD (Wedad et al., 2008; Abd Elmoneim et al., 2017). Germinated sorghums were reported to have higher protease activity and higher amino nitrogen as well as increased albumin and globulin fractions and a general decline in the kafirin fraction (Yi et al., 2017). However, Abd Elmoneim et al. (2017) argued that the effect of germination on the inaccessible proteins was minimal, while fermentation plays an essential role in modifying the protein aggregates to make them more accessible. In our study, germination/sprouting was shown to have a significant and positive effect on IVPD, with all test genotypes except Macia having almost as high IVPD as its uncooked version. Macia has vitreous endosperm (Figure 4-1) reported to be rich in enzyme resistant cross-linked polymeric protein (Ioerger et al., 2007). However, grain hardness alone does not seem to be responsible since other genotypes such as B503 are even more vitreous and perhaps as hard as Macia and yet had reasonably higher IVPD under all-grain processing treatments.

Sorghum proteins become less digestible when cooked in any form. The current result also showed that for all treatments, IVPD dropped upon cooking compared to their uncooked state and the unprocessed control but with significant differences between treatments and genotypes (Table 4-6). Cooked IVPD dropped the most in roasted samples and the least in sprouted samples. For the majority of the genotypes, cooked IVPD from sprouted samples was 91-92% as high as the unprocessed raw sample. This is the least drop in IVPD upon cooking of sorghum foods and sprouting appears to have significant potential in addressing the problem of protein digestibility in sorghum, especially for smallholder sorghum producers in Africa. Cooked IVPD in all other processing treatments was less than 70% of the unprocessed raw sample, with the most significant reduction occurring in roasted samples. The chemical changes brought about by the processing treatments to cause the difference in IVPD, cooked or uncooked, are unclear.

The impact of flour particle size on IVPD in cooked and uncooked states was significant (Table 4-2). Finer particle size flours and food samples had higher IVPD than the coarser samples (Table 4-4 and Table 4-6). This agrees with previous studies where smaller particle sizes, perhaps due to increased surface area for enzymatic action contributed to improved IVPD. This appears to be primarily because on starch than protein per se in that fine milling causes more disturbance to starch granules and make them more prone to attack by hydrolyzing enzymes which is believed to open the way for protease enzymes to act on freed protein bodies. Peralta-Contreras et al. (2013)

reported that coarser granules in sorghum resulted in a lower fermentation rate and slow degradation of starch. Depending on genotypes and processing treatments, particle size distribution of samples milled with similar mesh sizes seems to vary and is related to IVPD. Generally, flours carrying higher proportion of small particles tend to have higher IVPD compared to those with lesser proportion of smaller particles. In this study, the particle size of genotypes for the four processing treatments tend to vary with roasted samples having larger proportion of large size particles compared to the sprouted sample which carries relatively less proportion of the large particles (Figure 4-3).

Conclusion

This study showed that protein digestibility is a complex trait that is affected by multitudes of factors. Cultivar type and food processing methods do have a significant impact on protein availability. Genotypes with inherently improved PC and IVPD milled to appropriate particle sizes and pre-processed prior to making fermented bread can significantly improve protein availability from sorghum foods. This study demonstrated that sprouting/germination of grains prior to further processing to make fermented bread could remarkably increase protein digestibility by minimizing the negative impact of cooking on IVPD. The process also increased PC, but the nutritional impact of carbohydrate loss that resulted in higher relative PC needs to be determined. Sprouting entails little or no cost for processing and can be easily adopted by smallholder communities in Africa to improve protein nutrition. On the other hand, the semi-roasting of sorghum grains prior to cooking, a common practice used by women in Africa to prepare grains for milling, has a negative impact on protein digestibility. Decortication does not seem to affect IVPD in tannin-free sorghum and thus may not be important for improving the protein digestibility of sorghum foods.

Reference

Abd Elmoneim, O.E., R. Bernhardt, G. Cardone, A. Marti, S. Iametti, et al. 2017. Physicochemical properties of sorghum flour are selectively modified by combined germination-fermentation. J Food Sci Technol 54(10): 3307–3313.

Adejuwon, K.P., O.F. Osundahunsi, S.A. Akinola, M.O. Oluwamukomi, and M. Mwanza. 2020. Effect of Fermentation on Nutritional Quality, Growth and Hematological Parameters of Rats Fed Sorghum-Soybean-Orange flesh Sweet Potato Complementary Diet. Food Sci Nutr.

Adeyeye, S.A.O., A.O. Adebayo-Oyetoro, O.E. Fayemi, H.K. Tiamiyu, E.K. Oke, et al. 2019. Effect of co-fermentation on nutritional composition, anti-nutritional factors and acceptability of cookies from fermented sorghum (Sorghum bicolor) and soybeans (Glycine max) flour blends. Journal of Culinary Science & Technology 17(1): 59–74.

Afify, A.E.-M.M.M.R.R., H.S. El-Beltagi, S.M. El-Salam, A.A. Omran, S.M. Abd El-Salam, et al. 2012. Protein solubility, digestibility and fractionation after germination of sorghum varieties. PLoS One 7(2): e31154. doi: 10.1371/journal.pone.0031154.

Alvarez, M.M., E. Pérez-Carrillo, and S.O. Serna-Sald'ivar. 2010. Effect of decortication and protease treatment on the kinetics of liquefaction, saccharification, and ethanol production from sorghum. Journal of Chemical Technology & Biotechnology 85(8): 1122–1129.

Attuquayefio, W.D. 2014. Influence of processing parameters on eye size and elasticity of tef-based injera.

Barcelos, C.A., R.N. Maeda, G.J. V Betancur, and N. Pereira Jr. 2011. Ethanol production from sorghum grains [Sorghum bicolor (L.) Moench]: evaluation of the enzymatic hydrolysis and the hydrolysate fermentability. Brazilian Journal of chemical engineering 28(4): 597–604.

Belton, P.S., and J.R.N. Taylor. 2004. Sorghum and millets: protein sources for Africa. Trends Food Sci Technol 15(2): 94–98.

Chavan, U.D., J.K. Chavan, and S.S. Kadam. 1988. Effect of fermentation on soluble proteins and in vitro protein digestibility of sorghum, green gram and sorghum-green gram blends. J Food Sci 53(5): 1574–1575.

Correia, I., A. Nunes, A.S. Barros, and I. Delgadillo. 2010. Comparison of the effects induced by different processing methods on sorghum proteins. J Cereal Sci 51(1): 146–151.

Cremer, J.E., L. Liu, S.R. Bean, J.-B. Ohm, M. Tilley, et al. 2014. Impacts of kafirin allelic diversity, starch content, and protein digestibility on ethanol conversion efficiency in grain sorghum. Cereal Chem 91(3): 218–227.

Day, C.N., and R.O. Morawicki. 2016. Effects of fermentation by yeast and amylolytic lactic acid bacteria on grain sorghum protein content and digestibility. J Food Qual 2018.

Demeke, M., and F. Di Marcantonio. 2019. Analysis of incentives and disincentives for sorghum in Ethiopia. Gates Open Res 3(913): 913.

Derso, T., A. Tariku, G.A. Biks, and M.M. Wassie. 2017. Stunting, wasting and associated factors among children aged 6–24 months in Dabat health and demographic surveillance system site: A community based cross-sectional study in Ethiopia. BMC Pediatr 17(1): 96. doi: 10.1186/s12887-017-0848-2.

Dicko, M.H., H. Gruppen, O.C. Zouzouho, A.S. Traoré, W.J.H. Van Berkel, et al. 2006. Effects of germination on the activities of amylases and phenolic enzymes in sorghum varieties grouped according to food end-use properties. J Sci Food Agric 86(6): 953–963.

Duodu, K.G., A. Nunes, I. Delgadillo, M.L. Parker, E.N.C. Mills, et al. 2002. Effect of grain structure and cooking on sorghum and maize in vitro protein digestibility. J Cereal Sci 35(2): 161–174.

Duodu, K.G., J.R.N. Taylor, P.S. Belton, and B.R. Hamaker. 2003. Factors affecting sorghum protein digestibility. J Cereal Sci 38(2): 117–131.

Elkhalifa, A.E.O., R. Bernhardt, and others. 2010. Influence of grain germination on functional properties of sorghum flour. Food Chem 121(2): 387–392.

Emmambux, N.M., and J.R.N. Taylor. 2003. Sorghum kafirin interaction with various phenolic compounds. J Sci Food Agric 83(5): 402–407.

FAOSTAT. 2018. FAOSTAT. http://www.fao.org/faostat/en/#home.

Fox, J., and S. Weisberg. 2018. An R companion to applied regression. Sage publications.

Funke, I., and M.F. Melzig. 2005. Effect of different phenolic compounds on $α$-amylase activity: screening by microplate-reader based kinetic assay. Die Pharmazie-An International Journal of Pharmaceutical Sciences 60(10): 796–797.

Gali, B., and P.P. Rao. 2012. Regional analysis of household consumption of sorghum in major sorghum-producing and sorghum-consuming states in India. Food Secur 4(2): 209–217.

Gebre, A., P.S. Reddy, A. Mulugeta, Y. Sedik, and M. Kahssay. 2019. Prevalence of Malnutrition and Associated Factors among Under-Five Children in Pastoral Communities of Afar Regional State, Northeast Ethiopia: A Community-Based Cross-Sectional Study. J Nutr Metab 2019: 1–13. doi: 10.1155/2019/9187609.

Gebreegziabher, T., and N. Regassa. 2019. Ethiopia's high childhood undernutrition explained: analysis of the prevalence and key correlates based on recent nationally representative data. Public Health Nutr 22(11): 2099–2109. doi: 10.1017/S1368980019000569.

Gerald J., and Dorothy R. Friedman. 2019. Global Dietary Database. School of Nutrition Science and Policy at Tufts University.

Hothorn, T., F. Bretz, P. Westfall, R.M. Heiberger, A. Schuetzenmeister, et al. 2016. Package 'multcomp.' Simultaneous inference in general parametric models. Project for Statistical Computing, Vienna, Austria.

Ioerger, B., S.R. Bean, M.R. Tuinstra, J.F. Pedersen, J. Erpelding, et al. 2007. Characterization of polymeric proteins from vitreous and floury sorghum endosperm. J Agric Food Chem 55(25): 10232–10239.

Kumar, V., A.K. Sinha, H.P.S. Makkar, and K. Becker. 2010. Dietary roles of phytate and phytase in human nutrition: A review. Food Chem 120(4): 945–959.

Kumar, P.M.H., T.K. Virupaksha, and P.J. Vithayathil. 1979. SORGHUM PROTEINASE INHIBITORS: 2. Mode of Interaction with Serine Proteinases. Int J Pept Protein Res 13(2): 153–160.

Lenth, R., H. Singmann, J. Love, P. Buerkner, and M. Herve. 2019. Emmeans: Estimated marginal means, aka least-squares means. 2018; R package version 1.3. 1.

Maclean Jr, W.C., G.L. de RomaÑa, R.P. Placko, and G.G. Graham. 1981. Protein quality and digestibility of sorghum in preschool children: balance studies and plasma free amino acids. J Nutr 111(11): 1928–1936.

Mahasukhonthachat, K., P.A. Sopade, and M.J. Gidley. 2010. Kinetics of starch digestion in sorghum as affected by particle size. J Food Eng 96(1): 18–28.

Mella, O.N.O. 2011. Effects of malting and fermentation on the composition and functionality of sorghum flour.

Mertz, E.T., M.M. Hassen, C. Cairns-Whittern, A.W. Kirleis, L. Tu, et al. 1984. Pepsin digestibility of proteins in sorghum and other major cereals. Proceedings of the National Academy of Sciences 81(1): 1–2.

Nikmaram, N., S.Y. Leong, M. Koubaa, Z. Zhu, F.J. Barba, et al. 2017. Effect of extrusion on the anti-nutritional factors of food products: An overview. Food Control 79: 62–73.

Osman, M.A. 2004. Changes in sorghum enzyme inhibitors, phytic acid, tannins and in vitro protein digestibility occurring during Khamir (local bread) fermentation. Food Chem 88(1): 129–134.

Peralta-Contreras, M., C. Chuck-Hernandez, E. Perez-Carrillo, G. Bando-Carranza, M. Vera-Garcia, et al. 2013. Fate of free amino nitrogen during liquefaction and yeast fermentation of maize and sorghums differing in endosperm texture. Food and Bioproducts Processing 91(1): 46–53.

Pickerell, A.T.W. 1986. The influence of free alpha-amino nitrogen in sorghum beer fermentations. Journal of the Institute of Brewing 92(6): 568–571.

R Core Team. 2020. R: A Language and Environment for Statistical Computing.

Rashwan, A.K., H.A. Yones, N. Karim, E.M. Taha, and W. Chen. 2021. Potential processing technologies for developing sorghum-based food products: An update and comprehensive review. Trends Food Sci Technol.

Ratnavathi, C. V, and J. V Patil. 2013. Sorghum utilization as food. Nutrition \& Food Sciences 4(1): 1–8.

Ronda, V., C. Aruna, K. Visarada, and B.V. Bhat. 2019. Sorghum for animal feed. Breeding Sorghum for diverse end uses. Elsevier. p. 229–238

Rooney, L.W., and R.L. Pflugfelder. 1986. Factors affecting starch digestibility with special emphasis on sorghum and corn. J Anim Sci 63(5): 1607–1623.

Semba, R.D. 2016. The Rise and Fall of Protein Malnutrition in Global Health. Ann Nutr Metab 69(2): 79–88. doi: 10.1159/000449175.

Singh, A., S. Sharma, and B. Singh. 2017. Effect of germination time and temperature on the functionality and protein solubility of sorghum flour. J Cereal Sci 76: 131–139.

Subramanian, V., N.S. Rao, R. Jambunathan, D.S. Murty, and B.V.S. Reddy. 1995. The effect of malting on the extractability of proteins and its relationship to diastatic activity in sorghum. J Cereal Sci 21(3): 283–289.

Tadesse, B.T., A.B. Abera, A.T. Tefera, D. Muleta, Z.T. Alemu, et al. 2019. Molecular Characterization of Fermenting Yeast Species from Fermented Teff Dough during Preparation of Injera Using ITS DNA Sequence. Int J Food Sci 2019.

Taffesse, A.S., P. Dorosh, and S.A. Gemessa. 2013. 3 Crop production in Ethiopia: regional patterns and trends. Food and agriculture in Ethiopia. University of Pennsylvania Press. p. 53–83

Taylor, J.R.N. 1983. Effect of malting on the protein and free amino nitrogen composition of sorghum. J Sci Food Agric 34(8): 885–892.

Taylor, J., and J.R.N. Taylor. 2002. Alleviation of the adverse effect of cooking on sorghum protein digestibility through fermentation in traditional African porridges. Int J Food Sci Technol 37(2): 129–137.

Tesso, T., G. Ejeta, A. Chandrashekar, C.-P. Huang, A. Tandjung, et al. 2006. A novel modified endosperm texture in a mutant high-protein digestibility/high-lysine grain sorghum (Sorghum bicolor (L.) Moench). Cereal Chem 83(2): 194–201.

Warle, B., C. Riar, S. Gaikwad, and V. Mane. 2015. Effect of germination on nutritional quality of soybean (Glycine Max). Red 1(1.3).

Weaver, C.A., B.R. Hamaker, and J.D. Axtell. 1998. Discovery of grain sorghum germ plasm with high uncooked and cooked in vitro protein digestibilities. Cereal Chem 75(5): 665–670.

Wedad, H.A., A.H. El Tinay, A.I. Mustafa, and E.E. Babiker. 2008. Effect of fermentation, malt-pretreatment and cooking on antinutritional factors and protein digestibility of sorghum cultivars. Pak J Nutr 7(2): 335–341.

Weerasooriya, D.K., S.R. Bean, Y. Nugusu, B.P. Ioerger, and T.T. Tesso. 2018. The effect of genotype and traditional food processing methods on in-vitro protein digestibility and micronutrient profile of sorghum cooked products (J.J. Loor, editor). PLoS One 13(9): e0203005. doi: 10.1371/journal.pone.0203005.

Yetneberk, S., L.W. Rooney, and J.R.N. Taylor. 2005. Improving the quality of sorghum injera by decortication and compositing with tef. J Sci Food Agric 85(8): 1252–1258.

Yi, C., Y. Li, and J. Ping. 2017. Germination of sorghum grain results in significant changes in paste and texture properties. J Texture Stud 48(5): 386–391.

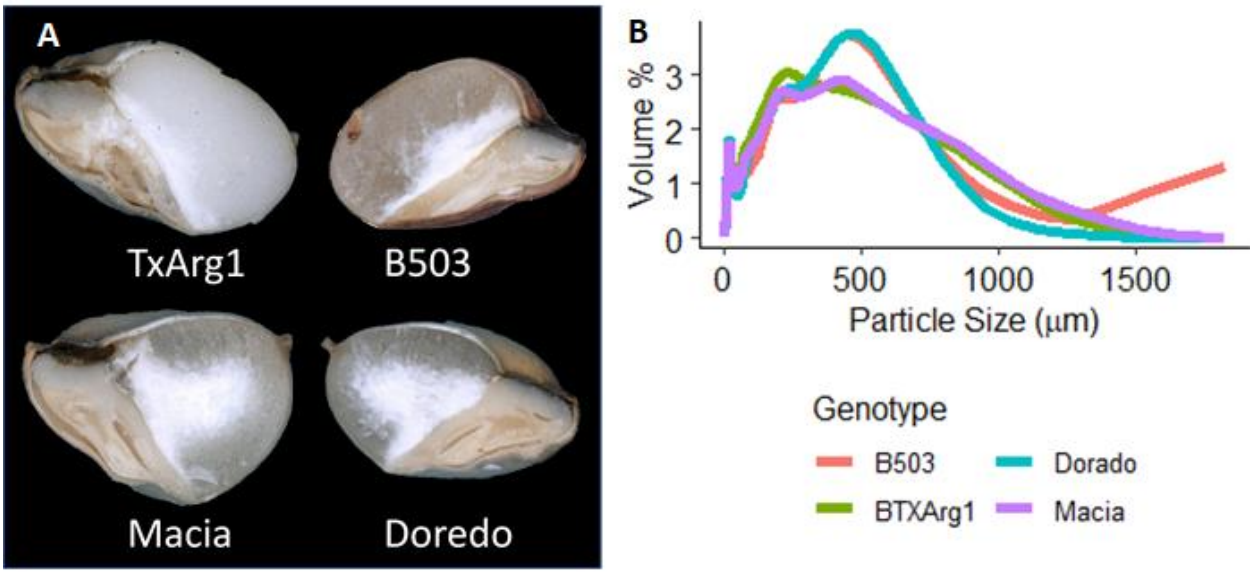


Figure 4-1 Physical properties of sorghum samples used in the study: (A) Endosperm vitreosity of the study genotypes (TxArg1 has waxy endosperm). (B) Particle size distribution of raw flour samples of the study genotypes milled using 2mm screen.

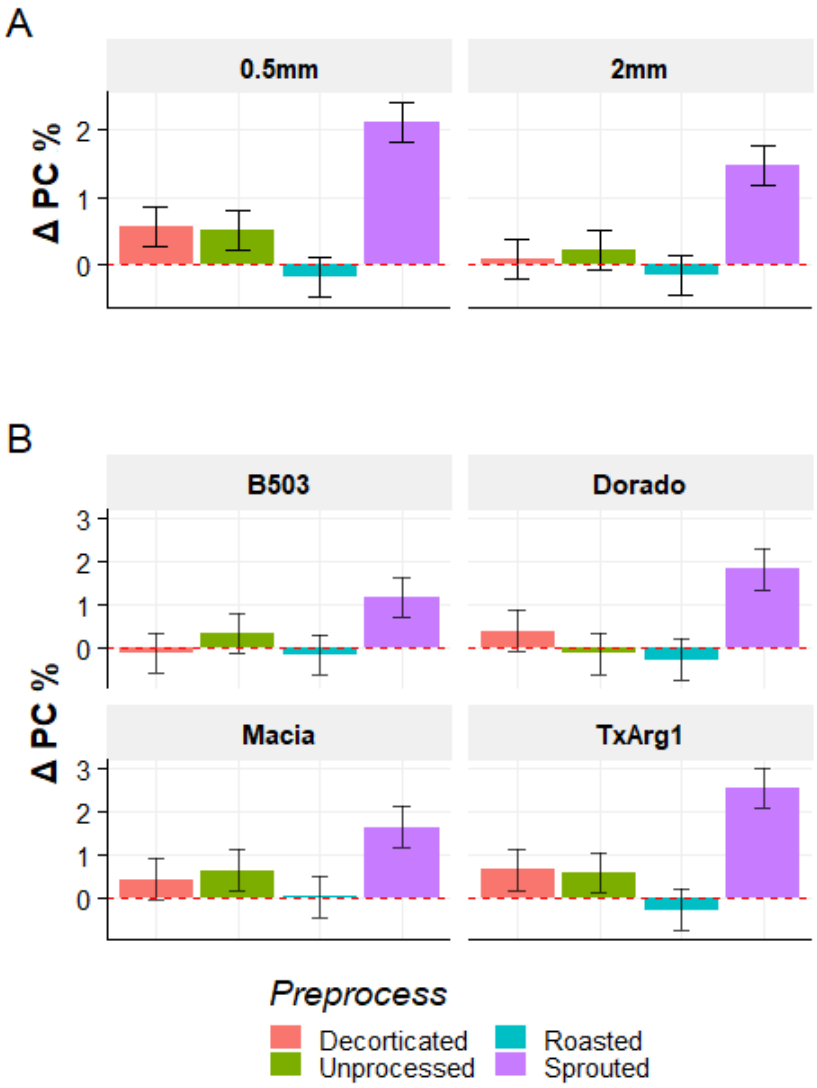


Figure 4-2 Estimated change in protein content (ΔPC) of fermented bread subjected to different pre-processing treatments as affected by particle size (A) and genotypes (B).

For the 0.5 mm screen, changes in protein content between uncooked and cooked samples were non-zero for all but the roasted method. However, for the courser 2 mm screen, only the sprouted samples showed non-zero ΔPC for all varieties.

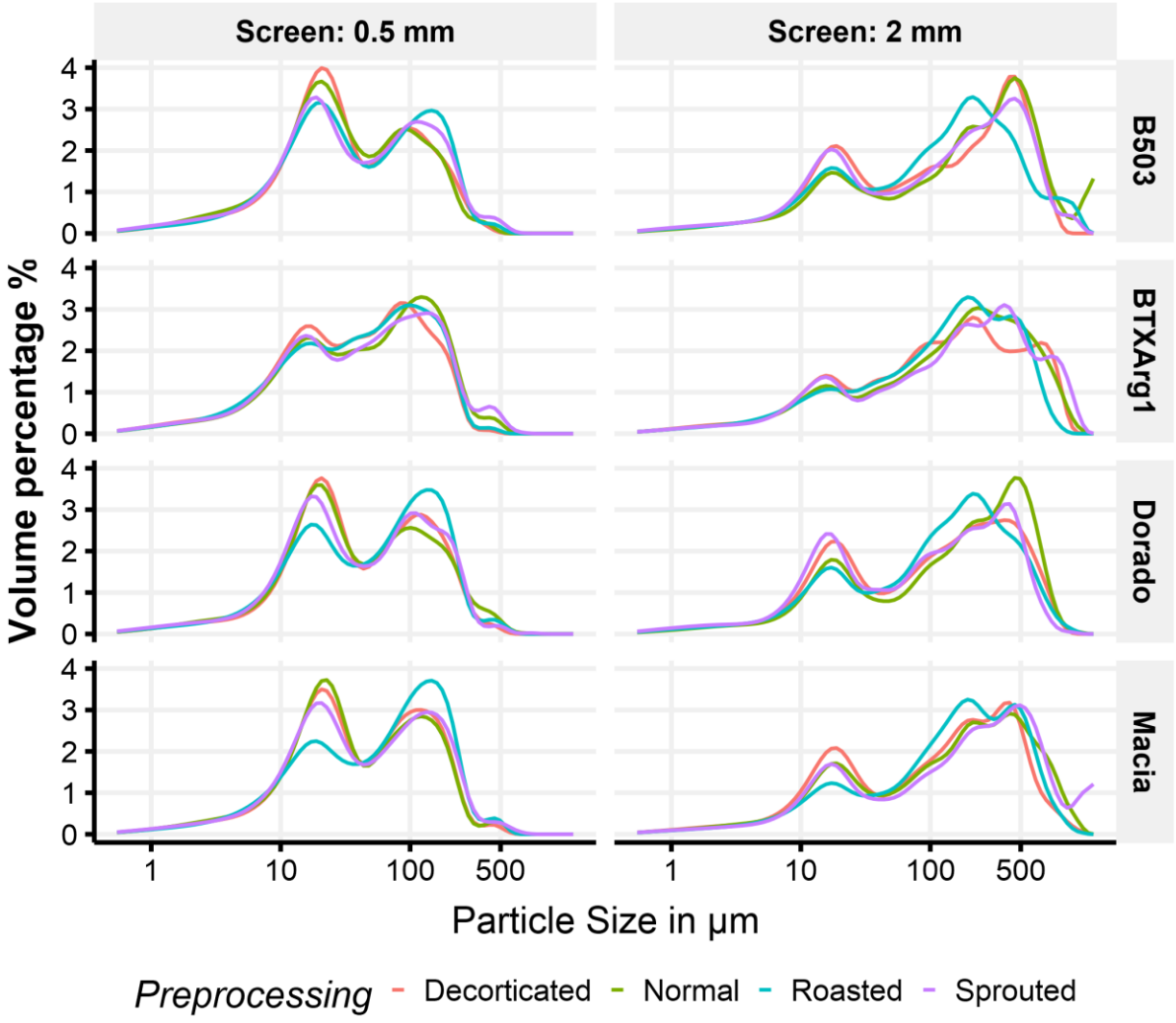


Figure 4-3 particle size distribution of different samples aggregated by variety, preprocessing method, and screen size.

The chart had two peaks. In the roasted samples, the left peak, which represented finer particles around 20 µm was underrepresented compared to most other pre-processing treatments.

Table 4-1 Physical and biochemical grain attributes the test sorghum genotypes estimated on a dry weight basis.

Genotype	Protein Content (%)	Amylose content (% of flour)	Vitreosity proportion (%)	Grain Color
B503	14.4 (± 0.09)	14.36 (± 0.55)	71.1 ± (11.25)	Red
TXArg1	13.7 (± 0.11)	0.37 (± 0.35)	Waxy endosperm	White
Dorado	13.1 (± 0.06)	14.26 (± 0.54)	57.3 (± 1.50)	White
Macia	12.0 (± 0.09)	14.57 (± 0.10)	63.2 (± 7.23)	White

Table 4-2 Analysis of variance on the effect of sorghum genotype, processing treatments, and screen size on in-vitro protein digestibility (IVPD), protein content (PC) and change in PC (∆PC) of cooked and uncooked samples.

Source of Variation	DF	IVPD (Raw)	PC (Raw)	IVPD (cooked)	PC (cooked)	∆PC
Replication	1	2.00	0.2405	0.07	1.46	1.64
Genotype (G)	3	2.25	438.15**	11.64**	205.24**	10.26**
Preprocess (P)	3	37.23**	16.15**	110.32**	210.57**	117.29**
Flour particle Size (S)	1	40.61**	3.30	6.33*	45.79**	19.43**
G x P	9	1.43	1.83	5.63**	6.10**	4.41**
G x S	3	0.65	0.72	1.32	2.62.	1.27
P x S	3	0.99	0.68	2.57	6.91**	3.62*
G x P x S	9	0.58	0.32	1.31	1.27	1.16
Error	31					

*, ** statistically significant at $P \leq 0.05$ and $P \leq 0.01$ levels of probability, respectively.

Table 4-3 The mean PC (%) of uncooked sorghum food samples as affected by processesing, genotype, and screen used.

	Genotypes				
Pre-processing treatments	B503	Dorado	Macia	TxArg1	Mean
2 mm screen size					
Decorticated	13.9	12.9	11.7	12.9	12.8 B
Unprocessed	14.4	13.1	12.0	13.8	13.3 A
Roasted	14.2	13.1	11.9	13.7	13.2 A
Sprouted	14.5	13.1	12.0	13.5	13.3 A
Mean	14.2 a	13.0 c	11.9 d	13.5 b	13.15 *ns
0.5 mm screen size					
Decorticated	14.1	12.9	11.8	13.1	13.0 B
Unprocessed	14.5	13.1	11.9	13.6	13.3 A
Roasted	14.5	13.1	11.9	13.7	13.3 A
Sprouted	14.6	13.4	12.1	13.5	13.4 A
Mean	14.43 a	13.13 c	11.93 d	13.48 b	13.25 *ns
Combined					
Decorticated	14.0	12.9	11.8	13.0	12.9 B
Unprocessed	14.4	13.1	12.0	13.7	13.3A
Roasted	14.3	13.2	11.9	13.7	13.3 A
Sprouted	14.5	11.8	12.0	13.5	13.3 A
Mean	14.3 a	13.1c	11.9 d	13.5 b	13.20

The mean PC of uncooked sorghum food samples as affected by cooking processes, genotype, and flour particle size. Means in the same column followed by the same uppercase letters are not significantly different at $P \leq 0.05$. Means in the same row followed by the same lowercase letters are not significantly different at $P \leq 0.05$). Bold faced means with (*) followed by different letters show a significant difference between screen size treatment means.

Table 4-4 In-vitro protein digestibility (IVPD) of uncooked sorghum food samples affected by pre-processesing, genotype, and flour particle size treatments.

	Genotypes				Mean
Pre-processing treatments	B503	Dorado	Macia	TxArg1	
2 mm screen size					
Decorticated	45.2	37.4	43.2	53.0	44.7 A
Unprocessed	36.2	37.4	44.3	48.8	41.7 A
Roasted	27.0	23.5	23.0	23.6	24.3 B
Sprouted	45.0	45.5	46.2	60.5	49.3 A
Mean	38.4	36.0	39.2	46.5	**40.0 *b**
0.5 mm screen size					
Decorticated	63.4	55.3	47	64.9	57.7 A
Unprocessed	58.7	59.0	60.5	59.6	59.4 A
Roasted	40.9	28.4	30.1	29.4	32.2 B
Sprouted	51.9	62.4	68.5	66.0	62.2 A
Mean	53.7	51.3	51.5	51.3	**52.9 *a**
Combined					
Decorticated	54.3 A	46.4 A	45.1 A	59.0 A	51.2 A
Unprocessed	47.4 AB	48.2 A	52.4 A	54.2 A	50.6 A
Roasted	33.9 B	26.0 B	26.5 B	26.5 B	28.2 B
Sprouted	48.4 AB	53.9 A	57.4 A	63.2 A	55.7 A
Mean	46.1	43.6	45.4	48.9	46.4

Means in the same column followed by the same uppercase letters are not significantly different at $P \leq 0.05$. The Means in the same row followed by the same lowercase letters are not significantly different at $P \leq 0.05$)—bold-faced means with (*) followed by different letters show a significant difference between screen size treatment means.

Table 4-5 The effect of preprocessing, genotype, and particle size on the PC cooked food samples.

	Genotypes				Mean
Pre-processing treatments	B503	Dorado	Macia	TxArg1	
2 mm particle size					
Decorticated	13.3	13.0	12.1	13.3	12.9 C
Unprocessed	14.5	13.1	12.4	14.1	13.5 B
Roasted	14.1	12.8	12.0	13.3	13.1 C
Sprouted	15.2	14.7	13.2	15.7	14.7 A
Mean	14.3 a	13.4 b	12.4 c	14.1 a	**13.6 *b**
0.5 mm particle size					
Decorticated	14.5	13.6	12.2	14.0	13.6 B
Unprocessed	15.1	12.8	12.8	14.5	13.8 B
Roasted	14.2	12.9	11.8	13.5	13.1 C
Sprouted	16.2	15.4	14.1	16.4	15.5 A
Mean	15.0 a	13.7 c	12.7 d	14.6 b	**14.0 *a**
Combined					
Decorticated	13.9 C.a	13.3 B b	12.2 B c	13.6 C ab	13.3 B
Unprocessed	14.8 B a	12.9 B b	12.6 B b	14.3 B a	13.6 B
Roasted	14.2 B a	12.8 B c	11.9 C d	13.4 C b	13.0 C
Sprouted	15.7 A a	15.1 A b	13.7 A c	16.0 A a	15.1 A
Mean	14.6 a	13.5 c	12.6 d	14.3 b	13.8

Means in the same column followed by the same uppercase letters are not significantly different at $P \leq 0.05$. Means in the same row followed by the same lowercase letters are not significantly different at $P \leq 0.05$). Bold-faced means with (*) followed by different letters show a significant difference between screen size treatment means.

Table 4-6 The effect of processing treatment, genotype, and screen size on IVPD of cooked sorghum food samples.

Pre-processing treatments	Genotypes				Mean
	B503	Dorado	Macia	TxArg1	
2 mm particle size					
Decorticated	28.5	24.9	24.3	30.2	27.0 B
Unprocessed	30.2	28.7	32.4	28.9	30.1 B
Roasted	20.7	12.3	10.4	11.9	13.8 C
Sprouted	44.4	46.0	32.5	44.9	41.9 A
Mean	30.9 a	28.0 ab	24.9 b	29.0 ab	**28.2 *b**
0.5 mm particle size					
Decorticated	39.3	32.4	22.8	36.0	32.6 B
Unprocessed	29.2	28.4	30.8	30.8	29.8 B
Roasted	25.3	13.5	19.6	16.2	18.7 C
Sprouted	43.2	41.8	27.7	54.1	41.7 A
Mean	34.3 a	29.0 ab	25.2 b	34.2 a	**30.7 *a**
Combined					
Decorticated	33.9 B a	28.6 B ab	23.6 AB b	33.1 B a	29.8 B
Unprocessed	29.7 BC a	28.6 B a	31.6 A a	29.9 B a	29.9 B
Roasted	23.0 C a	12.9 C b	15.0 B ab	14.0 C ab	16.2 C
Sprouted	43.8 A a	43.9 A a	30.1 A b	49.5 A a	41.8 A
Mean	32.6 a	28.5 bc	25.1 c	31.6 ab	29.4

Means in the same column followed by the same uppercase letters are not significantly different at $P \leq 0.05$. Means in the same row followed by the same lowercase letters are not significantly different at $P \leq 0.05$). Bold-faced means with (*) followed by different letters show a significant difference between screen size treatment means.

Table 4-7 Changes in in-vitro protein digestibility (IVPD) of cooked sorghum food products caused by processing treatments.

Genotype	Preprocess	Raw IVPD (%)	Cooked IVPD (%)	IVPD Change (%)	Cooked IVPD (%) relative to uncooked control
B503	Decorticated	54.3	33.9	-37.6	71.5
Dorado	Decorticated	46.4	28.6	-38.4	59.3
Macia	Decorticated	45.1	23.6	-47.7	45.0
TxArg1	Decorticated	59	33.1	-43.9	61.1
B503	Roasted	33.9	23	-32.2	48.5
Dorado	Roasted	26	12.9	-50.4	26.8
Macia	Roasted	26.5	15	-43.4	28.6
TxArg1	Roasted	26.5	14	-47.2	25.8
B503	Sprouted	48.4	43.8	-9.5	92.4
Dorado	Sprouted	53.9	43.9	-18.6	91.1
Macia	Sprouted	57.4	30.1	-47.6	57.4
TxArg1	Sprouted	63.2	49.5	-21.7	91.3
B503	Unprocessed	47.4	29.7	-37.3	62.7
Dorado	Unprocessed	48.2	28.6	-40.7	59.3
Macia	Unprocessed	52.4	31.6	-39.7	60.3
TxArg1	Unprocessed	54.2	29.9	-44.8	55.2

Appendix A - Supplementary Material Chapter 2

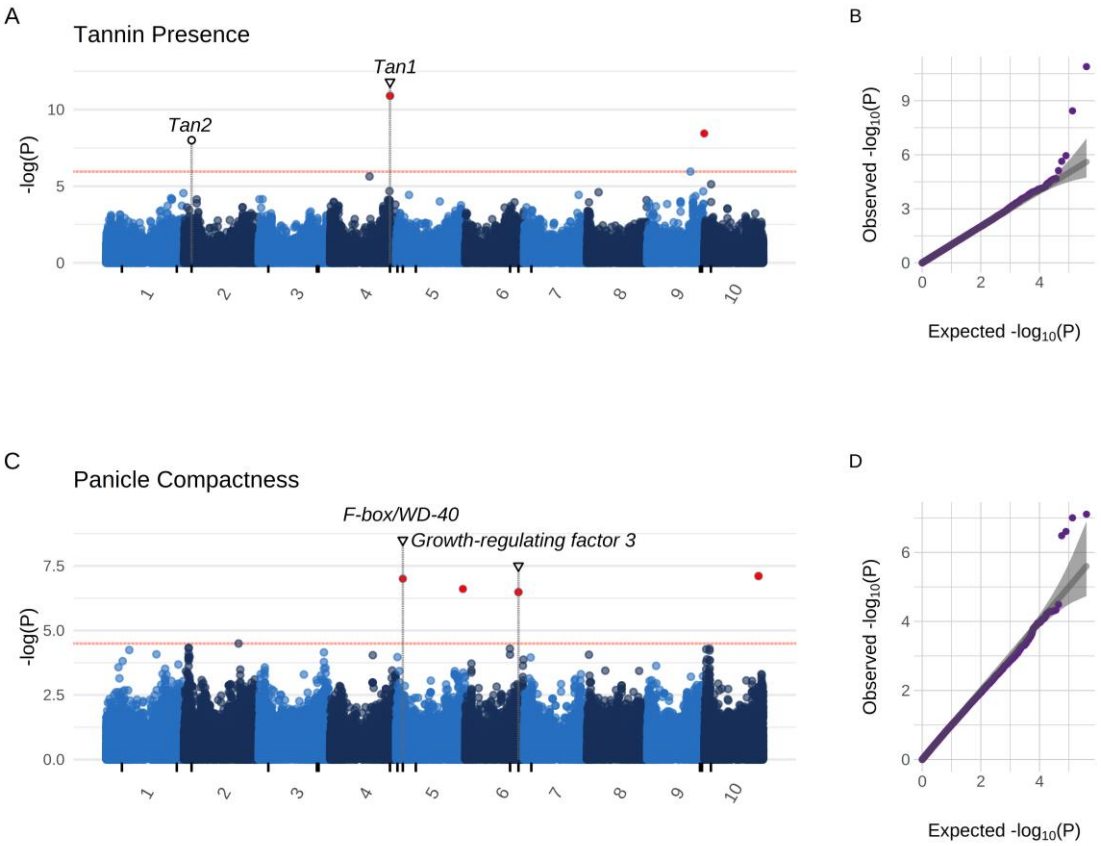


Figure A-1 Manhattan plot for Genome-wide association study using Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway (BLINK) and its associated quantile-quantile plot for Tannin presence (A and B,respectively), and Panicle compactness (C and D, respectively).

Red horizontal lines represent threshold at FDR correction (α =0.05). The vertical lines show some linked (<50 kbp) genes with associated SNPS.

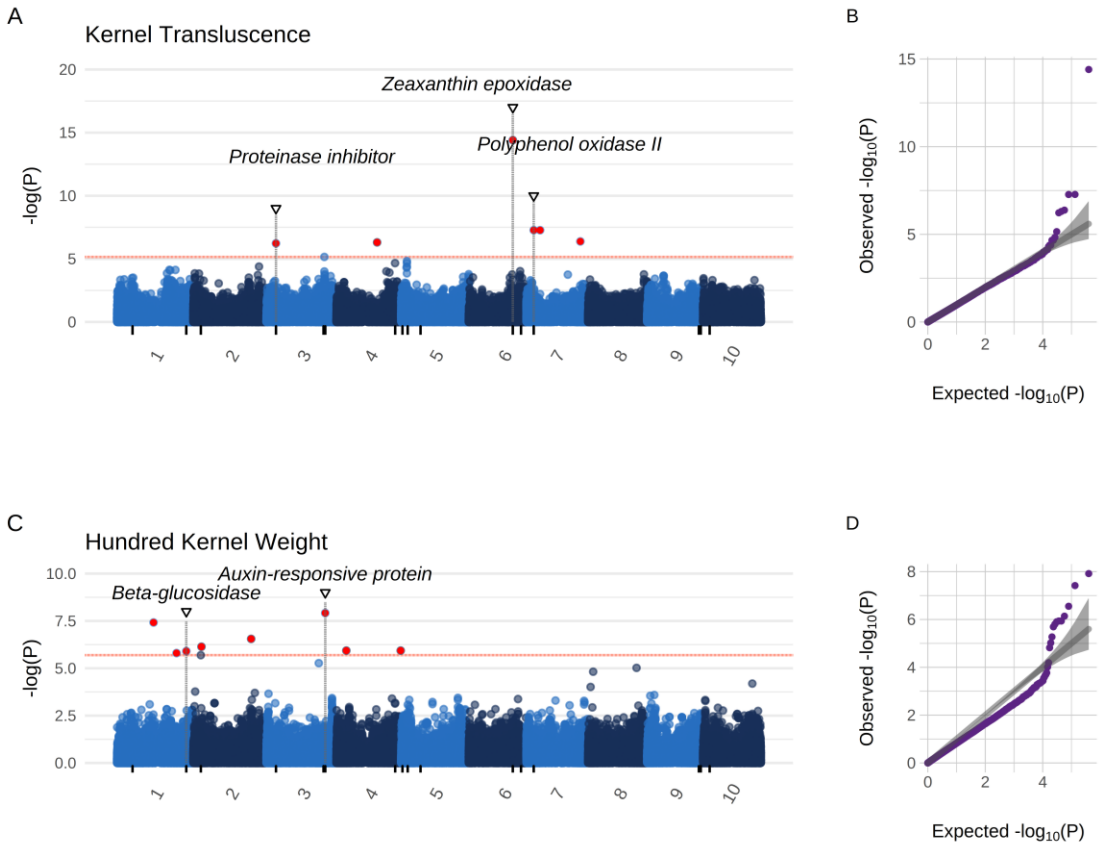


Figure A-2 Manhattan plot for Genome-wide association study and its respective quantile-quantile plot for kernel transluscence using BLINK (A and B) and for Hundred Kernel weight using Fixed and random model Circulating Probability Unification (FARMCPU) (C and D).

Red horizontal lines represent FDR adjusted P-value at α =0.05. Red dots represent SNPs singficantly associated with the precipitation variables. The vertical lines show some of the linked (<50 kbp) genes with singnificantly associated SNPS.

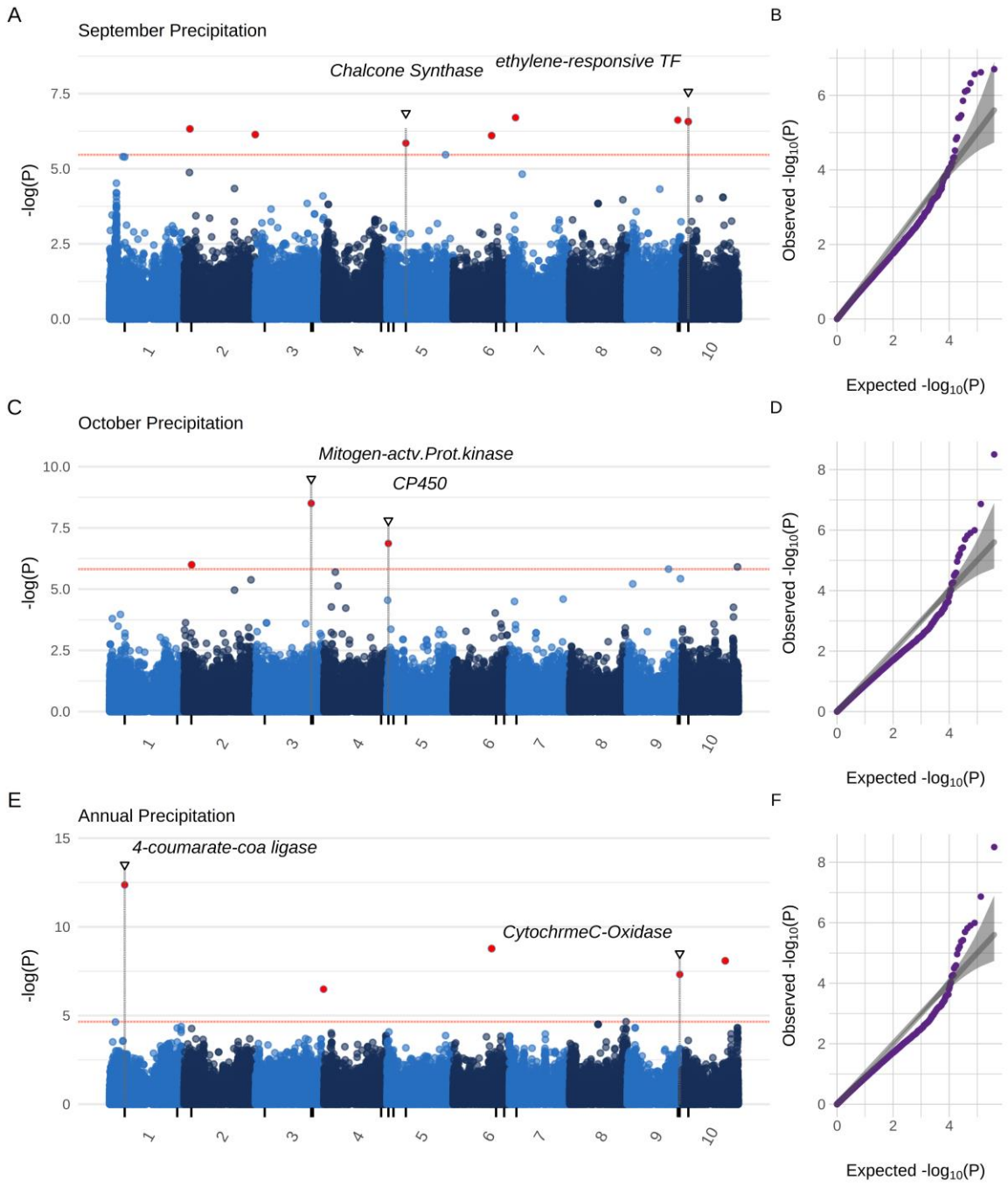


Figure A-3 Genome-environment association study (using FARMCPU) for precipitation variables (A, C) and respective quantile-quantile (Q-Q) plots (B, D).

Red horizontal lines represent FDR adjusted P-value at α =0.05. Red dots represent SNPs singficantly associated with the precipitation variables. The vertical lines show some of the linked (<50 kbp) genes with singnificantly associated SNPS.

Table A-1 List of Priori genes related to sorghum storage proteins, anthocyanin synthesis and starch properties extracted from NCBI database.

NCBI Gene Symbol	Ensemble Gene ID	Function	Source	Chromosome	Position Start	Position End
LOC8062726	SORBI_3002G211700	glutelin-2 gamma kafirin protein gamma-kafirin preprotein	NCBI	2	60423442	60424313
LOC8068650	SORBI_3009G007100	glutelin type-B 5	NCBI	9	639511	641454
LOC8068296	SORBI_3005G188800	kafirin PGK1 Kafirin PSK8 kafirin preprotein	NCBI	5	67366389	67367331
LOC8079123	SORBI_3005G192900	kafirin PSKR2-like alpha kafirin	NCBI	5	67651628	67652562
LOC8074403	SORBI_3005G184800	zein-alpha Z4 19kD-like alpha kafirin B3	NCBI	5	66970093	66971018
LOC8072915	SORBI_3010G136100	10 kDa prolamin delta-kafirin seed storage protein delta kafirin truncated delta-kafirin	NCBI	10	20530171	20530521
LOC8068297	SORBI_3005G189000	kafirin PSKR2-like	NCBI	5	67373729	67374693
LOC8066724	SORBI_3005G184700	zein-alpha A20 19kD-like alpha kafirin B1 23 kDa alpha-kafirin	NCBI	5	66965527	66966456
LOC8066721	SORBI_3005G184400	kafirin PSKR2-like	NCBI	5	66923278	66924099
LOC8062726	SORBI_3002G211700	glutelin-2 gamma kafirin protein gamma-kafirin preprotein	NCBI	2	60423442	60424313
LOC110435706	SORBI_3005G184500	kafirin PSKR2-like	NCBI	5	66926726	66927556
LOC110435321	SORBI_3005G193100	kafirin PSKR2-like alpha kafirin	NCBI	5	67654898	67655764
LOC110435320	SORBI_3005G192700	kafirin PSKR2 alpha kafirin putative kafirin preprotein	NCBI	5	67638681	67639585
LOC110435318	SORBI_3005G193140	kafirin PSKR2-like alpha kafirin	NCBI	5	67658193	67658999
LOC110435317	SORBI_3005G193180	kafirin PSKR2-like alpha kafirin	NCBI	5	67661336	67662286
LOC110429512	SORBI_3005G193000	kafirin PSKR2-like alpha kafirin hdhl 22-kDa alpha kafirin	NCBI	5	67648393	67649339

NCBI Gene Symbol	Ensemble Gene ID	Function	Source	Chromosome	Position Start	Position End
LOC8079125	SORBI_3005G193260	kafirin PSKR2-like alpha kafirin	NCBI	5	67667809	67668707
LOC8079124	SORBI_3005G193220	kafirin PSKR2-like alpha kafirin	NCBI	5	67664574	67665522
LOC8079122	SORBI_3005G192801	kafirin PSKR2-like alpha kafirin	NCBI	5	67641925	67642823
LOC8079121	SORBI_3005G192901	kafirin PSKR2-like alpha kafirin	NCBI	5	67645163	67646057
LOC8074401	SORBI_3005G184600	kafirin PSKR2-like	NCBI	5	66931562	66932272
LOC8066726	SORBI_3005G185400	kafirin PSKR2-like	NCBI	5	67014861	67015858
LOC8065278	SORBI_3009G001600	zein-beta beta-kafirin truncated beta-kafirin	NCBI	9	166827	167614
LOC8060747	SORBI_3002G055000	regulatory protein opaque-2 opaque 2 protein	NCBI	2	5254989	5257807
LOC8060745	SORBI_3002G054800	regulatory protein opaque-2	NCBI	2	5243140	5247362
LOC8057484	SORBI_3006G108832	protein FLOURY 1	NCBI	6	47817214	47818167
LOC8056591	SORBI_3006G175700	anthocyanin regulatory R-S protein	NCBI	6	53102701	53111029
LOC8056590	SORBI_3006G175500	anthocyanin regulatory R-S protein	NCBI	6	53062306	53080184
LOC8081981	SORBI_3004G328800	anthocyanin 5-aromatic acyltransferase	NCBI	4	66305270	66307053
LOC8080419	SORBI_3001G340900	anthocyanin regulatory C1 protein	NCBI	1	62819065	62821694
LOC8079334	SORBI_3010G178700	anthocyanin 5-aromatic acyltransferase	NCBI	10	51681001	51683061
LOC8065008	SORBI_3010G269700	anthocyanin regulatory R-S protein	NCBI	10	60405194	60407174
LOC8063845	SORBI_3002G139200	malonyl-coenzyme:anthocyanin 5-O-glucoside-6'''-O-malonyltransferase	NCBI	2	21729220	21731045
LOC8058211	SORBI_3003G232900	anthocyanin 3'-O-beta-glucosyltransferase	NCBI	3	57192514	57194484
LOC8073018	SORBI_3006G076900	anthocyanin regulatory R-S protein myc-like regulatory R gene product	NCBI	6	44126383	44140389

NCBI Gene Symbol	Ensemble Gene ID	Function	Source	Chromosome	Position Start	Position End
LOC8056589	SORBI_3006G175200	anthocyanin regulatory R-S protein myc-like regulatory R gene product	NCBI[1]	6	53044178	53049377
LOC8055322	SORBI_3004G348700	malonyl-coenzyme A:anthocyanin 3-O-glucoside-6"-O-malonyltransferase	NCBI	4	67766000	67768720
LOC8076027	SORBI_3004G280800	WD40 repeat *(TRANSPARENT TESTA GLABRA 1) Tan1*	NCBI	4	62315396	62318779
LOC8069098	SORBI_3002G076600	basic helix-loop-helix protein A (*Tan2*)	NCBI	2	7975937	7985221
LOC8068390	SORBI_3010G022600	Granule-bound starch synthase 1 (*Waxy*)	NCBI	10	1860965	1865278

Priori genes , not listed here but related with grain weight and panicle compactness can be found from (Olatoye et al., 2018; Wang et al., 2020).

Tao, Y., Zhao, X., Wang, X., Hathorn, A., Hunt, C., Cruickshank, A. W., van Oosterom, E. J., Godwin, I. D., Mace, E. S., & Jordan, D. R. (2020). Large-scale GWAS in sorghum reveals common genetic control of grain size among cereals. Plant Biotechnology Journal, 18(4), 1093–1105.

Wang, J., Hu, Z., Upadhyaya, H. D., & Morris, G. P. (2020). Genomic signatures of seed mass adaptation to global precipitation gradients in sorghum. Heredity, 124(1), 108–121.

[1] NCBI gene database: Gene. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2020 04 20]. Available from: https://www.ncbi.nlm.nih.gov/gene/