A comprehensive phenotypic and genomic characterization of Ethiopian sorghum germplasm defines core collection and reveals rich genetic potential in adaptive traits

Gezahegn Girma1 | Habte Nida1 | Alemu Tirmessa2 | Dagnachew Lule3 | Tamirat Bejiga2 | Amare Seyoum2 | Moges Mekonen2 | Amare Nega2 | Kebede Dessalegn3 | Chemeda Birhanu3 | Alemnesh Bekele4 | Adane Gebreyohannes2 | Getachew Ayana2 | Tesfaye Tesso5 | Gebisa Ejeta6 | Tesfaye Mengiste1

1 Department of Botany and Plant Pathology, Purdue University, West Lafayette, IN 47907, USA
2 Ethiopian Institute of Agricultural Research, P.O. Box 2003, Addis Ababa, Ethiopia
3 Oromia Agricultural Research Institute, P.O. Box 81265, Addis Ababa, Ethiopia
4 Haramaya University, P.O. Box 138, Dire Dawa, Ethiopia
5 Department of Agronomy, Kansas State University, 3007 Throckmorton PSC 1712 Claffin Road, Manhattan, KS 66506, USA
6 Department of Agronomy, Purdue University, West Lafayette, IN 47907, USA

Correspondence
Tesfaye Mengiste, Department of Botany and Plant Pathology, Purdue University, West Lafayette, IN 47907, USA.
Email: tmengist@purdue.edu

Funding information
United States Agency for International Development

Abstract
Understanding population genetic structure and diversity of a crop is essential in designing selection strategies in plant breeding. About 2010 Ethiopian sorghum accessions were phenotyped for different traits at multiple locations. A subset of the collection, 1628 accessions, predominantly landraces, some improved varieties, and inbred lines were genotyped by sequencing. Phenotypic data revealed association of important traits with different sorghum growing agro-climatic regions, high genetic diversity and the presence of rare natural variation in the Ethiopian sorghum germplasm. Subsequent genotypic analysis determined optimum number of sub-populations, distinct cluster groups and ancestries of each sorghum accessions. To improve utilization of germplasm, a core subset of 387 lines were selected following posteriori grouping of genotypes based on cluster groups obtained through GBS analysis followed by stratified random sampling using quantitative traits. In order to evaluate how well this new sorghum and millet innovation lab (SMIL) collection from Ethiopia is represented within the largest world sorghum collection at United States Department of Agriculture - National Plant Germplasm System (USDA-NPGS) and the sorghum association panel (SAP), comparisons were conducted based on SNP data. The SMIL collection displayed high genetic diversity with some redundancy with the USDA-NPGS germplasm but SAP showed clear distinction. Furthermore,
1 | INTRODUCTION

_Sorghum bicolor_ (L.) Moench is a staple food crop with particular importance to subsistence farming in semi-arid tropical regions of the world (Leff, Ramankutty, & Foley, 2004). These areas are known for erratic rainfall and moisture deficit, often not suitable for other cereal crops. Drought tolerance of sorghum is partly attributed to its C4 physiology, a physiological mechanism that makes plants more efficient while energy making in hot, sunny and dry climates, and in turn, results in higher productivity. The C4 photosynthetic carbon cycle has evolved as an adaptation to high light intensities, high temperatures, and drought, which allows them to dominate grassland floras and biomass production in the warmer climates of the tropical and subtropical regions (Edwards, Osborne, Strömberg, & Smith, 2010). Several other adaptive mechanisms such as thick leaf wax, deep root system and stay green trait allow its tolerance under extreme drought condition (Tari, Laskay, Takacs, & Poor, 2013).

Evolved in one of the major centers of crop diversity in the world with wide ranging environmental conditions (Plant Genetic Resources Center, 1995), Ethiopian sorghums harbor significant diversity for important traits such as cold (Singh, 1985), and drought tolerance (Adugna, 2014), disease resistance (grain mold and various leaf diseases) (Nida et al., 2019; Weerasooriya et al., 2016) and nutritional profile (Rhodes et al., 2017). Of the known main sorghum races, i.e. bicolor, guinea, caudatum, dura and kafir, all but kafir are grown in Ethiopia (Engels & Hawkes, 1991). As a result, today’s sorghum crop possesses broad ecological adaptation where it is very commonly grown in diverse climates, from the dry lowland areas of the semi-arid tropics receiving very little annual precipitation, to high rainfall and humid regions, as well as in the more temperate regions of the world.

According to data regularly compiled by the Food and Agricultural Organization (FAO) of the United Nations, from 1994 to 2018, of the top ten sorghum producing nations, more than 67% of the global sorghum production is contributed by developing countries (FAOSTAT, 2020). However, sorghum productivity in developing countries in general and sub-Saharan Africa in particular, remains low as its production is pre-dominantly with seeds of low yielding landrace germplasm cultivated in soils that are inherently nutrient deficient, and in regions with erratic rainfall and low/high temperature stresses. Use of modern technologies, inorganic fertilizer and improved seed of sorghum varieties among small holder farmers of this region continues to be low (Mundia, Secchi, Akamani, & Wang, 2019). Moreover, pest and disease limit sorghum productivity (Kumari et al., 2016; Wortmann et al., 2009). Broad genetic diversity is essential for providing the gene pool necessary for the development and production of new seed varieties of crops with badly needed agriculturally important traits. Unfortunately, vast variation of gene pools is often harbored in raw landrace germplasm or their wild and related species of native crops that evolved in Vavilov centers of origin and diversity of crop plants such as Ethiopia (Hummer & Hancock, 2015). Hence, both the collection and assembly, as well as the phenotypic and genotypic characterization of such a collection, are key steps in a well thought out and careful crop improvement and utilization program. Large-scale germplasm characterization with next-generation sequencing based genotyping techniques such as genotyping-by-sequencing (GBS), facilitate exploration and discovery of novel traits to expedite genetic gain during crop improvement process by enabling selection of a diverse genetic pool. Several efforts have been made to characterize Ethiopian sorghum germplasm collection ranging from 80 to 974 accessions (Adugna, 2014; Ayana & Bekele, 2000; Ayana, Bryngelson, & Bekele, 2000; Desmae, Jordan, & Godwin, 2016; Weerasooriya et al., 2016). However, these studies were limited in sample size and lacked representation of the large geographic coverage of Ethiopian sorghum collections that are held at several sites. The Ethiopian Biodiversity Institute (EBI) holds a total of 9249 sorghum accessions (IBC, 2007). A collection of Ethiopian sorghums is also maintained at different genebanks around the world, including over 2000 accessions conserved at the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) (Upadhyaya, Sharma, Dwivedi, & Singh, 2014) and about 7400 accessions preserved at genebanks in the National Plant Germplasm System (NPGS) of the United States Department of Agriculture (USDA) (Cuevas,
The ever decreasing cost and high throughput sequencing allowed genotyping of many individuals, increasing the power of genome-wide association studies (GWAS). GWAS have enabled identification of loci and/or genes underlying complex traits, such as yield (Li et al., 2019), disease resistance (Wang et al., 2012) and other quantitative traits (Girma et al., 2019). It is known that quantitative traits are highly influenced by environment and genetics. Genome-environment association (GEA) studies based on GWAS models have created additional avenue for exploring the genetic basis of plant adaptation to abiotic stresses (Lopez-Hernandez & Cortes, 2020; Torre, Wilhite, & Neale, 2019).

In this study, phenotypic data from field trials, genotypic data generated through genotyping by sequencing, coupled with agro-climatic data of sorghum landraces representing Ethiopian sorghum germplasm collection were used to conduct comprehensive germplasm analyses. A representative core subset that captures the diversity of the Ethiopian sorghum germplasm was developed as a critical resource for sorghum breeding, genomic and genetic studies. Furthermore, we implemented genome-environment association (GEA) studies to explore the genetic basis of adaptive traits. Overall, we provide a better insight into the Ethiopian sorghum germplasm, which is important in the world sorghum collection for its features relative to other collections.

2 MATERIALS AND METHODS

2.1 Plant materials and passport data

A total of about 2010 Ethiopian sorghum germplasm entries comprising landrace, improved varieties and inbred lines were included in this study (Supplemental Table S1). These germplasms were assembled from several sources by the sorghum research project at Ethiopian Institute of Agricultural Research (EIAR), as part of a United States Agency for International Development (USAID) funded sorghum and millet innovation lab (SMIL) project, and the collection is hereafter described as the SMIL collection. Most of the accessions representing landraces were derived from a larger collection of sorghum germplasm maintained at Ethiopian Biodiversity Institute (EBI). Representative samples were drawn from a list of more than 9000 accessions by taking major sorghum geographic regions, occurrence of different stressors, local production, and utilization practices into consideration. The original passport data included only accession and locality names. Passport data was further enriched by including mean annual temperature (°C), annual precipitation (mm), altitude (m asl) and agro-climatic information following commonly used agro-climatic and Köppen-Geiger climate classification (Kottek, Jurgen, Christoph, Bruno, & Franz, 2006) for each accession based on available locality name (Supplemental Table S1). The traditional
agro-climatic classification is largely based on mean annual precipitation and altitude.

### 2.2 Test sites and phenotyping

All accessions (N = 2010) were planted at Bako (9°08’ N latitude, 37°03’ E longitude, 1642 m asl altitude, 19.7 °C mean annual temperature, 1281 mm mean annual rainfall), Arsi Negelle (7°21’ N latitude, 38°42’ E longitude, 1951 m asl altitude, 17.7 °C mean annual temperature, 915 mm mean annual rainfall) and Haramaya (9°24’ N latitude, 42°01’ E longitude, 2047 m asl altitude, 17.9 °C mean annual temperature, 799 mm mean annual rainfall). Each accession was planted in a single row of 3 m length with 20 cm distance between plants and 80 cm between plots during beginning of May to mid-May of 2015 and 2016 at Arsi Negelle, Haramaya and Bako. Fertilizer was applied at the rate of 100 kg/ha DAP at planting and 50 kg/ha urea at knee height. Pure lines in each accession were obtained by selecting and selfing five single heads representing the most frequent genotype within a row. The homogeneity was confirmed by planting individuals for at least one more season. Five plants representing each accession were tagged and data on sixteen morphological traits (Supplementary Table S2) were collected following the standard descriptors (IBPGR and ICRISAT, 1993) used in sorghum germplasm characterization. Six quantitative traits (days to flowering, days to physiological maturity, plant height, seed number, grain yield and thousand kernel weight) were collected from the three locations. SPAD and all categorical traits were collected at Arsi Negelle (Supplementary Table S2).

### 2.3 Genotyping-by-sequencing

A total of 1628 accessions including 1585 landrace, 32 cultivars and 11 inbred lines were genotyped by sequencing (Supplemental Table S1). Of which 40 individuals were genotyped redundantly as intentional duplicates to determine any genotyping error and exclude identical accessions; where 37 were in duplicate, two in triplicate and one seventeen times. DNA from dried tissue of week-old seedlings after germination was isolated following a modified CTAB procedure (Mace, Buhariwalla, Buhariwalla, & Crouch, 2003). A total of seventeen 96-plex ApeKI reduced representation libraries were constructed and genotyped following genotyping by sequencing (GBS) procedure (Elshire et al., 2011) at University of Wisconsin, Biotechnology Center. SNP datasets are deposited in Purdue University Research Repository https://purr.purdue.edu/publications/3189/2 (https://doi.org/10.4231/EZDX-XT41).

In addition, other datasets (Supplemental Table S3) previously generated through similar genotyping procedure were included with the current SMIL dataset for further analysis. The materials represent 285 SAP lines among genotypes used in previous studies (Morris et al., 2013) and a core subset of 380 accessions representing 5% of the 7400 Ethiopian sorghum germplasms maintained at USDA-NPGS as part of the largest global sorghum collection previously described (Cuevas et al., 2017).

### 2.4 Data analysis

#### 2.4.1 Phenotypic data analysis

Descriptive statistics (range, mean and standard error) were computed for seven quantitative traits by categorizing accessions into different agro-climatic groups, commonly used classification of sorghum growing areas in Ethiopia. Pearson correlation coefficients and confidence interval of the correlation coefficient at 95% were computed across quantitative traits using mean values from all the locations. However, as SPAD data was collected only at Arsi Negelle its correlation coefficient was also computed with other quantitative traits obtained from the same site. Frequency distribution for categorical data across all accessions were visualized using histogram.

#### 2.4.2 GBS variant calling and quality control

Single nucleotide polymorphism (SNP) calling was performed using the DiscoverySNPCallerPluginV2 of the Tassel GBS pipeline v5 (Glaubitz et al., 2014) following sequence read alignment to the reference genome of Sorghum bicolor version 3.1.1 (McCormick et al., 2018) using Burrows-Wheeler Alignment (BWA) tool (Li & Durbin, 2009). The SNP calling analysis was done for SMIL collection separately, and in unison with other datasets representing SAP lines and Ethiopian sorghum core subset at USDA-NPGS. Quality control of SNP data was performed considering percent missing individual, missing marker, minor allele frequency and excess heterozygosity with PLINK version 1.90 (Purcell et al., 2007).

#### 2.4.3 Population genetic structure and genetic diversity analysis

Pairwise genetic distance was calculated among accessions genotyped in duplicate. Any variation between two individuals from the same genotype was considered due
to genotyping error. Accordingly, genetic distance threshold was set for declaring two accessions as identical. To describe population structure of the SMIL collection, different complementary approaches were implemented. Hierarchical clustering and neighbor joining tree were made by calculating pairwise genetic distance (identity-by-state, IBS) in PLINK version 1.90 (Purcell et al., 2007). A Ward’s minimum variance hierarchical cluster dendrogram was then built from IBS matrix using an R (R Core Team, 2017) package called Analyses of Phylogenetics and Evolution (APE) (Paradis, Claude, & Strimmer, 2004). Secondly, model-based maximum likelihood estimation of individual ancestries from multi-locus SNP genotype datasets using ADMIXTURE method (Alexander & Lange, 2011) which assumes linkage equilibrium among loci and Hardy-Weinberg equilibrium within ancestral populations was used to identify ancestries of each sorghum accession. The admixture analysis was performed for different K (number of sub-populations) varying from 2 to 20. The most appropriate K-value was selected after considering 10-fold cross-validations whereby the best K exhibits low cross-validation error compared to other K values.

To further understand patterns of genetic relatedness across accessions, principal components analysis (PCA) was conducted using ‘prcomp’ function in R (R Core Team, 2017). This was further looked in terms of partitioning of genomic variation due to improvement status (landrace/released landrace/improved variety/inbred line) and germplasm source (SMIL/USDA-NPGS/SAP) across sorghum accessions. Genetic diversity parameters which include inbreeding coefficient (FIS), observed heterozygosity (Ho), expected heterozygosity (He), minor allele frequency (MAF) and polymorphic information content (PIC) across sorghum accessions were generated using TASSEL (Glaußit et al., 2014) and PLINK version 1.90 (Purcell et al., 2007).

2.4.5 Core subset selection and validation

To further improve utilization of germplasm, a core subset was selected following posteriori grouping of genotypes (1941 accessions representing landrace) based on genetic cluster group obtained through GBS analysis followed by stratified random sampling using quantitative traits. The optimum number of K with lower cross validation error value determined through admixture analysis was considered as the most appropriate number of sub-populations. All genotyped accessions were classified into one of the sub-populations and remaining accessions with no genotype data were treated separately as an independent group. Stratified selection of core subset as previously suggested (Franco, Crossa, Taba, & Shands, 2005) was made using a ccChooser package (Studnicki & Debski, 2012) in R software (R Core Team, 2017). The core selection was validated by comparing differences in range, mean and variances of entire and core collections for quantitative traits. Shannon-Weaver diversity index (H) (Shannon, C. E. & Weaver, 1949) was used to check frequency distribution for categorical traits. Chi-square ($\chi^2$) probability was used to determine frequency distribution based on agro-climatic group. Furthermore, neighbor joining tree for all genotyped landrace accessions was generated to visualize distribution of the core selected across entire collection.

2.4.6 Genome-environment association (GEA) analysis

Agro-climatic information including altitude, mean annual temperature and precipitation for each locality, representing the collection area of 1425 sorghum landrace accessions and robust SNP markers that passed through quality control from GBS data were used. An association analysis was performed with various GWAS models including general linear model (GLM) with genotype data corrected for population structure, mixed linear model (MLM) with genotype data corrected for both population structure and kinship (Zhang et al., 2010) and Bayesian-information and linkage-disequilibrium iteratively nested keyway (BLINK), an enhanced version of the fixed and random model circulating probability unification (FarmCPU) method (Huang et al., 2018), that substitutes kinship with a set of markers associated with the casual genes. GWAS was implemented in the GAPIT package (Lipka et al., 2012) with the three different models in R software (R Core Team, 2017). Log Q–Q plots of p-values were examined to determine how well the models accounted for population structure. Significant
FIGURE 1  Agro-climatic classification of Ethiopian sorghum accessions in the SMIL collection. (a) traditional climate classification, (b) Köppen-Geiger climate classification, (c) annual mean precipitation and (d) annual mean temperature across a range of altitudes of sorghum growing areas. NA represents number of accessions that could not be assigned to any of the agro-climatic groups due to missing information associations were determined for each trait by adjusting the raw p-value using the conservative Bonferroni error rate control method. Manhattan and Q–Q plots were visualized using the qqman R package (Turner, 2014). All the significant SNP markers were mapped onto the most recent version of Sorghum bicolor v3.1.1 (McCormick et al., 2018) genome in Phytozome v12.1 (Goodstein et al., 2012).

3 | RESULTS

3.1 | Geographic description of the study materials

The current collection was the largest studied both in terms of sample size and representation across sorghum growing agro-ecologies of Ethiopia; local steppe climate to warm and temperate climate with elevation ranging from 356–3162 m asl, mean annual temperature from 11.6–28.7°C and mean annual precipitation from 144–1955 mm (Figure 1; Supplemental Table S1). Accessions from highland areas with altitude above 1900 m asl represented the largest collection followed by dry lowland, wet lowland and intermediate altitudes. Likewise, grouping according to Köppen-Geiger climate classification indicated large number of accessions from tropical climate followed by warm and temperate climate with the least number of accessions from local steppe climate. The weather data also showed insignificant but a positive correlation ($r = 0.06$) between mean annual precipitation and altitude. Whereas, strong negative correlation ($r = -0.96$) were observed between mean annual temperature and altitude.

3.2 | Phenotypic diversity across SMIL collection

Ethiopian sorghum germplasm, currently studied under SMIL collection, represents diverse set of materials based on both quantitative and qualitative traits. Qualitative data collected in the study were summarized and presented in histograms (Figure 2). All quantitative traits measured were grouped and summarized based on agro-climatic information (Supplemental Table S4). The highest mean plant height of 331.33 cm was observed in germplasm from wet intermediate altitude with shorter plant types from the dry lowland areas. Likewise, both days to flowering and maturity were longer in the wet intermediate altitude accessions and shorter in dry lowland materials. The highest mean number of seeds per panicle was also observed among accessions from wet intermediate altitude with the least number of seeds recorded in dry lowland materials. However, thousand grain weight and the SPAD (soil plant analysis development) values were higher among germplasm from dry areas and lower in materials from wet intermediate altitude. Grain yield per panicle was highest in the dry intermediate altitude accessions and least
in wet lowland types. Overall, plant height ranged from a fairly dwarf type of about 100 cm to as tall as 500 cm, flowering from 76 to 157 days, maturity period ranging from 136 to 208 days, thousand grain weight (TGW) from 7 to 52 gm, grain yield per panicle (GY) from 0.1 to 213 gm, seed number from 20 to 8070 per head and SPAD values from 19.70 to about 70. Overall, Pearson correlation analysis among quantitative traits revealed significant association, except for non-significant and negative association between SPAD score and seed number per panicle (Supplemental Table S5). SPAD also showed significant negative correlation with DTF \( r = -0.16, P \leq .01 \), DTM \( r = -0.39, P \leq .01 \) and PHT \( r = -0.45, P \leq .01 \). However, the correlation between SPAD and TGW \( r = 0.24, P \leq .01 \) and GY \( r = 0.08, P \leq .01 \) was significant and positive.

### 3.3 Summary statistics of SNP markers

A total of 879,407 SNP markers were initially discovered across 1628 individuals genotyped within SML collection. Overall, 46% of the SNPs called in the entire accessions had minor allele frequency (MAF) <0.05. Quality control of SNP data based on individuals with <20% missing and markers with <40% missing and MAF >0.05 produced 72,190 robust SNP markers with 1605 individuals passing filtering criteria. The number of SNPs was the lowest on chromosome 7 and the highest on chromosome 1. Both MAF and polymorphic information content (PIC) were the highest on chromosome 9 and lowest on chromosome 6. The highest and lowest level of heterozygosity was observed on chromosome 5 and chromosome 6, respectively (Supplemental Figure S1a). Overall, high-density
SNP markers were observed toward the tails across all chromosomes with almost no SNPs mostly in centromeric regions (Supplemental Figure S1b). Variant calling based on combined dataset of 2198 individuals (1541 from SMIL landrace collection, 285 SAP and 380 USDA-NPGS) resulted in large missing SNP markers across SAP population. This is likely due to improvement in sequence quality over time and batch effect. Hence, we implemented k nearest neighbors’ algorithm in Tassel (Glaubitz et al., 2014) for imputation of missing data. SNP calling for 2198 individuals initially discovered 646,413 SNPs. Quality control of SNP data based on individuals with <20% missing and markers with <60% missing and MAF >0.05 produced 106,330 SNP markers. SNP markers discovered following imputation were used for diversity analysis to compare USDA-NPGS and SAP lines with SMIL collection.

3.4 | Population genetic structure across SMIL collection

Following genetic distance analysis among duplicate accessions, a threshold value of 0.075 was established (Figure 3). Subsequently, a total of 1553 genetically distinct accessions (1501 landraces, 32 cultivars, 11 inbred lines and nine genotypes not assigned to any of these groups) were identified and used for further analysis (Supplemental Table S6). Analysis on pairwise distance-based hierarchical clustering identified distinct cluster groups based on 72,190 SNP markers (Figure 4a). Model-based maximum likelihood estimation of individual ancestries from multi-locus SNP genotype datasets using ADMIXTURE identified ancestries of each sorghum accessions. Admixture analysis also revealed K = 12 (Figure 4; Supplemental Figure S2; Supplemental Table S7), as the most appropriate number of sub-populations and showed good correspondence with the clustering pattern obtained by hierarchical cluster tree. PCA applied to genotype data to compare association between diversity and agroclimatic conditions of collection area explained differences among accessions. The first two principal components (PCs) explained 49.45% of the total variation with accessions from wet highland, wet lowland and wet intermediate altitudes showing some tendency to form distinct cluster group (Supplemental Figure S3).

3.5 | Genetic relationship and ancestry among cultivars and inbred lines

A total of 43 genotypes representing 20 improved cultivars, 11 inbred lines and 12 released landraces were also analyzed for diversity and maximum likelihood estimation of individual ancestries (Figure 4c; Supplemental Table S8; Supplemental Figures S4 and S5). Two groups of improved cultivars including Melkam, Meko and Misikir among the first and Dagim, Birimash and IS9302 in the second group shared more than 99% ancestry within group. Almost all improved cultivars except Baji have backgrounds of
mainly these two groups. Baji is a highly admixed cultivar with genetic background from multiple genotypes and significant contribution from released landraces including Adukara and Chiro/Chelenko/ETS2752 group. Interestingly, improved cultivars, Melkam/Meko/Misikir group, shared more than 99% ancestry with a commonly cultivated released landrace, Gambella1107. Released landraces mostly have distinct background but some of them shared high genetic ancestry. Chiro, Chelenko and ETS2752 shared more than 94%; Muyra-1 and Muyra-2 shared more than 99% and Dano and Gemedi shared more than 97% ancestry.

3.6 Sorghum core subset

A core subset of 387 accessions representing 20% of the SMIL collection at EIAR were selected (Supplemental Table S1). Shannon-Weaver diversity index across the two groups for categorical traits also confirmed that the core set captured maximum diversity and represented the entire collection (Supplemental Table S9). Frequency comparison based on Chi-square for entire and core collection showed no significant difference at \( P < 0.05 \) between the entire and core collection (Supplemental Table S10). Likewise, core evaluation based on range, mean and variance comparison of the entire and core collection showed no significant differences at \( P < 0.05 \) for all the seven morphological traits (Supplemental Table S11). Neighbor joining tree for 1501 genotyped accessions with 72,190 SNPs clearly indicated an even distribution of the core selected lines across all groups (Supplemental Figure S6).

3.7 Comparison of genetic diversity, population structure and genome-wide linkage disequilibrium

To compare and evaluate diversity captured in the SMIL collection, we have included SAP and a core subset representing 5% of 7400 Ethiopian sorghum germplasm within global collection maintained at USDA-NPGS. The SMIL collection in the current study represented higher diversity with mean genetic distance value of about 0.30 followed by 0.28 in the USDA-NPGS core collection and the lowest genetic distance of 0.27 was observed in the SAP. Histogram showing genetic distance frequency distribution across genotypes indicated maximum values above 0.43 within SMIL collection and less in the other collections (Supplemental Figure S7). Summary of genetic diversity statistics per taxa showed higher mean proportion of heterozygosity in SMIL followed by USDA-NPGS core subset. Likewise, per marker statistics revealed high mean PIC, MAF, expected (He) and observed (Ho) heterozygosity among SMIL population, followed by USDA-NPGS core subset. Conversely, higher mean inbreeding coefficient (\( F_{IS} \)) were observed in the SAP population with the least in SMIL collection (Supplemental Table S12).

Principal component analysis (PCA) clearly revealed distinction of the SAP with the first two principal components explaining 71.49% of the total variation (Figure 5a).
FIGURE 5  Population genetic structure across sorghum accessions obtained from different source. (a) PCA of 2198 individuals with PC axis calculated using 106,330 SNPs. The proportion of the variance explained by each PC is shown in parentheses. Dots with different color represents SAP (red), SMIL (green) and USDA-NPGS (blue), (b) individual ancestry estimated based on ADMIXTURE analysis for different K values. Each individual accession is represented as a single vertical line partitioned into segments corresponding to the inferred membership. The lower bar corresponds to germplasm source

The USDA-NPGS core subset collection showed higher diversity than SAP but was less diverse than the SMIL collection. Model-based maximum likelihood estimation of individual ancestries from multi-locus SNP genotype datasets using ADMIXTURE identified ancestries of each sorghum accessions (Figure 5b; Supplemental Table S13). A high ancestry membership coefficient (≥ 0.60) were observed at K = 11 among 68.95% of the genotypes within SMIL, 53.24% within USDA-NPGS and 47.92% in the SAP population, with remaining respective percentages showing admixed accessions.

Genome-wide LD measured as $r^2$ (the squared correlation coefficients among alleles at two adjacent SNP markers) equals to 0.20 within 6 kb, 7 kb and 9 kb on average in SAP, USDA-NPGS and SMIL populations, respectively (Figure 6). The LD decays to $r^2 < 0.1$ within 45 kb in the SAP, 55 kb in USDA-NPGS and 65 kb in the SMIL.

3.8  Genome-environment association study for agro-climatic parameters

Genome-environment association (GEA) analysis was conducted using genotype data of 1425 sorghum landraces with their respective agro-climatic information including mean annual temperature, mean annual precipitation and altitude. The same genotypic data has been used in our recent GWAS analysis across multiple traits which validated previous observations particularly for plant height and pericarp color (Girma et al., 2019). In addition, the distribution of the current agro-climatic dataset showed approximately normal distribution (Supplemental Figure S8). Three GWAS models were implemented and
evaluation of Q-Q plot across different methods in our GEA analysis indicated BLINK as the most appropriate model, reducing model-based errors, compared to GLM and MLM (Figure 7d, e and f). GEA results based on BLINK were therefore considered for further candidate gene identification following error rate control of raw p-values with Bonferroni correction.

A total of 15 different SNPs with significant association, with Bonferroni correction at an error rate of ≤ 0.05, were identified for the three agro-climatic parameters (Figure 7; Supplemental Table S14). Subsequently, 23 unique candidate genes at or closest to the significant SNPs for all agro-climatic parameters were identified.

A total of seven SNPs on chromosome 1, 5, 8, 9 and 10 were significantly associated (P ≤ 4.27 × 10⁻¹⁷) with variation in temperature (Supplemental Table S14). In total, 16 candidate loci at or within 10 kb vicinity of the significant SNPs were identified. Among these, one represents non-genic region, four were with in coding sequences for genes with no annotation or known protein domains and the remaining 11 represented genes with annotation. Of these, at least five were described to have biological function in plant adaption to temperature variation. The top most significantly associated SNP (S8_60346825), on chromosome 8, was located 2 kb from Srp40 (Serine/Arginine-rich proteins) which was previously reported to contribute to plant-specific forms of developmental regulation and environmental adaptation (Califice, Baurain, Hanikenne, & Motte, 2012). Four other candidate genes encoding F-box proteins, Auxin response factor 22, Pyrophosphate-energized vacuolar membrane proton pump (EC 3.6.1.1), and Ferritin 2 proteins were reported to mediate auxin action in biotic and abiotic stress responses, regulation of salt and drought stress tolerance and protection against oxidative stress (An et al., 2019; Bouzroud et al., 2018; Briat et al., 2010; Fukuda, Chiba, Maeda, Nakamura, & Maeshima, 2004).

Three SNPs in the vicinity of genic regions were associated with variation in altitude. A single SNP (S8_60346825)
closely linked (within 2 kb) to Sobic.008G169200 which encodes a protein weakly similar to suppressor protein Srp40 and Sobic.008G169000, encoding an F-box protein were significant for plant adaptation to altitudinal variation. In addition, Serine/Threonine-protein kinase wnk with no lysine-related which is known to respond differentially under various abiotic stresses including cold, heat, salt and drought (Kumar, Rao, Biswas, & Sinha, 2011) carried SNPs associated with temperature variation. Interestingly, this gene was also reported for its possible role in regulation of flowering time (Wang et al., 2008). Three of the six regions represent uncharacterized proteins with no functional annotation domain.

GEA analysis also identified 13 regions with six SNPs significantly associated (P ≤ 1.54 × 10−8) to variation in mean annual precipitation. Of which, nine represented transcripts either lacking annotation domain or uncharacterized putative protein. The remaining four candidate regions are annotated to carry a Myb3 transcription factor, Malate dehydrogenase, Phosphoribulokinase/Uridine kinase and small GTP-binding protein, were reported for their roles in adaptation to a various environment (Supplemental Table S14). These genes were previously described to control response to various biotic and abiotic stresses (Agarwal et al., 2008; Ambawat, Sharma, Yadav, & Yadav, 2013; Imran, Tang, & Liu, 2016; Janiak, Kwasniewski, Sowa, & Gajek, 2018).

4 | DISCUSSION

4.1 | Ethiopian sorghum genetic diversity and potential implications

Previous reports (Cuevas et al., 2017; Plant Genetic Resources Center, 1995) have described the Ethiopian sorghum as the most diverse crop distributed over a wide range of ecological variation in the country, ranging between 400 and 3000 m asl. The SMIL collection originally obtained from altitude ranges of 356–3162 m asl is, therefore, representative and captured variation in terms of both agro-climatic and trait variation. All qualitative traits we have described in this particular study have captured all descriptor categories for the traits (IBPGR and ICRISAT, 1993). Plausibly, high agro-ecological variability and existence of almost all sorghum race types explains the underlying diversity (Engels & Hawkes, 1991).

Of the different sorghum growing agro-ecological regions, wet intermediate altitude areas represent accessions with taller mean plant height, longer maturity period and high number of seeds per head compared to dry, wet lowland, and highlands (Supplemental Table S4). The wetter regions mostly represent the western parts of the country that is generally fertile, warm with high rainfall which is favorable for sorghum growth. However, the enormous potential of wet lowlands has not yet been exploited mainly due to high disease prevalence (Mengistu, Shimelis, Laing, & Lule, 2019) and soil acidity (Abebe, 2007).

As an important crop in several semi-arid regions of the world, observed diversity in traits associated with tolerance to moisture stress suggest the potential of this collection to serve as an important source for development of resilient varieties adapted to diverse environments. Our results show large diversity for morphological traits (plant height, days to flowering and maturity), physiological traits (stay green and chlorophyll concentration) and agronomic traits such as grain yield, thousand kernel weight and seed number (Figure 2; Supplemental Table S4) that may have relations with drought tolerance. In an earlier study based on sorghum association panel, grain yield and yield components such as number of grains and grain weight per panicle were reported to be associated with drought tolerance (Mutava, Prasad, Tuinstra, Kofoid, & Yu, 2011). Likewise, a significant correlation (P ≤ .01) was observed between grain yield and yield components (seed number and thousand kernel weight) described in the current study (Supplemental Table S5). Stay-green or delayed senescence in sorghum is an important trait associated with terminal drought tolerance and shows positive correlation with resistance to charcoal rot (Borrell, Mullet, George-jaeggli, & Van Oosterom, 2014; Duncan, Bockholt, & Miller, 1981). Here, we report that a significant number of accessions displayed the stay green trait (Figure 2). SPAD is a chlorophyll meter used as alternative measurement tool for leaf greenness used for easy and quick assessments of genotype tolerance to moisture stress (Mutava et al., 2011; Silva, Jifon, Da Silva, & Sharma, 2007). In the current study, SPAD values were higher among materials from the dry areas compared with materials from all other agro-climatic areas (Supplemental Table S4). Further correlation analysis among quantitative traits revealed positive association except for the negative correlation between SPAD with days to flowering, days to maturity, plant height and seed number. A similar study in barley indicated the negative correlation between days to heading and SPAD and positive correlation between grain yield and SPAD values (Bort, Araus, Hazzam, Grando, & Salvatore, 1998).

Ethiopian sorghum germplasm also harbors rare natural variants in the taxa. In this study, three double seeded genotypes were observed in the current SMIL collection (Figure 2). While their general impact on yield is not yet known, they clearly contribute to certain yield components such as number of kernels per panicle. A recent study identified ethyl methane sulfonate (EMS)-induced multiseeded (msd) mutants that result in full spikelet fertility, and
doubling grain number per panicle. And significance of doubled seeded genotypes was further demonstrated using a single base mutation which led to an increase in grain number by about 200% in sorghum (Gladman et al., 2019).

### 4.2 Genetic diversity and linkage disequilibrium

Molecular data demonstrated high diversity across the SMIL collection. A recent report on genomic characterization of a core subset of the Ethiopia sorghum within USDA-NPGS, representing 17% of the world largest collection were described as diverse both phenotypically and genetically (Cuevas et al., 2017). Here we detected even higher diversity among the new SMIL collection than both USDA-NPGS core subset and SAP. The SAP represented lower diversity and formed a distinct group. The uniqueness of the SAP from both Sudanese and Nigerien sorghum collection were similarly reported in recent studies (Cuevas & Prom, 2020; Maina et al., 2018).

The genome-wide LD decay within 7–9 kb for the landrace collection for USDA-NPGS and SMIL is in line with previous estimates, 7.3 kb at $r^2 = 0.2$ among 3909 diverse sorghum landraces (Hu, Olatoye, Marla, & Morris, 2019). The LD decay to $r^2 = 0.2$ was slightly faster in SAP which could be due to the recombination process through which the materials have undergone during the conversion program (Stephens et al., 1967).

### 4.3 Improved sorghum varieties from Ethiopia represent a narrow genetic diversity

PCA and unrooted neighbor-joining tree, both revealed reduced genetic diversity across improved sorghum varieties including inbred lines in the pipeline and released landraces (Supplemental Figures S4 and S5). Moreover, almost half of the varieties had ancestry membership coefficient with likelihood of more than 0.90 (Figure 4c). This has revealed narrow genetic base among majority of the cultivars developed in the breeding program with two genotypes (Melkam/Meko/Misikir and Dagim/Birimash/IS9302) representing the genetic backgrounds of all cultivars. The improved cultivars Melkam, Meko and Misikir were introduced from ICRISAT, evaluated for local adaptation and eventually released for production in the dry lowlands of Ethiopia. The high ancestry (99%) shared between a commonly cultivated released landrace, Gambella1107, with Melkam/Meko/Misikir group suggests that Gambella1107 was likely used as parental source. Gambella1107 was selected from ‘zera zera’, a subgroup of the race caudatum, common in Ethiopian-Sudan border (Rai, Murty, Andrews, & Bramel-Cox, 1999). Most modern cultivars in Ethiopia are expected to have a background from the ‘zera zera’ sorghum landraces. In addition to low diversity, most of these varieties have been developed for dry lowland climate, also suggesting lack of breeding effort for the wet lowland areas. This observation implies that even in countries like Ethiopia that harbor wide genetic variability for the species, genetic uniformity arising from recycling of parental sources can be real. This explicitly can pose risks both to crop production due to genetic vulnerability and also drag progress in improvement due to lack of genetic variability. It is important that future cultivar improvement efforts utilize information from the current and other studies to maximize variability in the breeding parents while at the same time maintaining acceptable threshold of target traits.

### 4.4 Core collection

Sorghum core collections have been established across genebanks, particularly at ICRI SAT and USDA-NPGS where the largest number of landraces are being maintained. Subsets of reduced sizes have been selected following different sampling strategies and stratification of the entire collection based on agro-morphological traits, response to photoperiod and passport data (Cuevas et al., 2017; Grenier & Hamon, 2001). However, the effect of environment on agro-morphological traits and limitation of passport information, especially for old collections, hinders our ability to adequately capture the genetic diversity of an entire collection within the core subset. In the current study, the use of genotyping by sequencing coupled with agro-morphological traits enabled the selection of a core collection representing the broader Ethiopian sorghum landrace germplasm.

Evaluation of diversity across accessions represented in the core subset through different methods (Supplemental Tables S9 to S11; Supplemental Figure S6) suggested maximum possible variation captured within 20% of the total collection. The newly developed core subset will, therefore, provide an easy access to efficient sorghum genetic resource utilization as the manageable size will enable easier multi-location evaluation. The selected core will also provide guidance for further collection missions to acquire new accessions. The availability of molecular profile in addition to representative samples of the large germplasm is also important for selection of genetically diverse parents for trait introgression and making crosses to generate mapping populations for different traits of interest.
4.5 Genome-environment association analysis revealed genetic basis of abiotic stress tolerance in sorghum

We have identified several loci and candidate genes underlying plant adaptation to variable environments that could be causal or at LD with causal genes. Most of these genes were described to control adaptation to various abiotic factors such as cold, heat, salt and drought tolerance and protection against oxidative stress (Supplemental Table S14). Comparable genome-wide association study on heat tolerance of sorghum developing leaves revealed several candidate genes associated to biological pathways involved in plant stress responses (Chen et al., 2017). Of the candidate genes suggested for various biological function, the $Srp40$ and $Myb3$ were implicated in plant developmental regulation (Ambawat et al., 2013; Califice et al., 2012). Our results also revealed two candidate genes, $F$-box domain and $Srp40$, which were significantly associated to plant adaptation across environments with both annual mean temperature and altitude data. This may not be surprising as temperature is linearly associated with altitude (Figure 1d) and are considered complementary to each other. Our study also revealed a candidate gene, Serine/Threonine-protein kinase wnk with no lysine-related, reported for possible roles in regulation of flowering time (Wang et al., 2008). The roles of environmental factors such as day length, temperature, drought and salinity in addition to endogenous genetic components determining flowering time across plant species including sorghum have been reported (Brunet & Larson-rabin, 2012; Cho, Yoon, & An, 2017; Cuevas et al., 2016). It has also been possible to exploit potential of day length-neutral sorghum genotypes of tropical origin that flower early in the temperate summer as it provides long days (Cuevas et al., 2016).

5 CONCLUSIONS

In this report, we demonstrated high diversity of the SMIL collection both in terms of phenotypic and genetic diversity parameters, population structure and LD. The faster LD decay within 9 kb across large and diverse landrace collection of sorghum will improve GWAS efficiency with high marker density. Narrow genetic diversity among the widely grown improved cultivars in Ethiopia, not representing the diversity in the landrace collections, displays underutilization of the existing genetic potential. This can also be considered risky due to increased recurrence of abiotic stresses and the changing pest and pathogen virulence as well as pathogen population dynamics. The genetic diversity and/or potential of these resources to grow across diverse environments has yet to be explored. Thus, it is recommended that breeders utilize diverse parental lines to capture novel genes or variants of existing genes/alleles for economic traits by taking advantage of the huge diversity in the Ethiopian sorghum germplasm. The developed core subset, representing the total diversity within SMIL collection, will provide easy access to germplasm by facilitating phenotyping across multiple environments, future genomics studies and, ultimately, selection of genotypes with desirable traits. Candidate loci identified with the GEA will have potential utilization for germplasm identification and sorghum breeding for stress prone area following further validation. The core collection needs further detailed characterization and cataloging of additional traits for effective and comprehensive utilization in crop improvement.

ACKNOWLEDGMENTS

This study is made possible through funding by the Feed the Future Innovation Lab for Collaborative Research on Sorghum and Millet through grants from American People provided to the United States Agency for International Development (USAID) under cooperative agreement number AID-OAA-A-13-00047. The contents are the sole responsibility of the authors and do not necessarily reflect the views of USAID or the United States Government.

AUTHOR CONTRIBUTIONS

T.T., G.E. and T.M. conceived the project idea. T.B., A.S., A.N., C.B., M.M., A.G., D.L, K.D., G.A. and A.B. performed field experiments. G.G. analyzed data and wrote the paper. H.N., A.T., D.L., T.T., G. E. and T.M. revised the manuscript. T.M. supervised the study. All authors read and approved the final version of the manuscript.

CONFLICT OF INTEREST DISCLOSURE

The authors declare no conflicts of interest.

ORCID

Gezahegn Girma https://orcid.org/0000-0001-9702-3521
Tesfaye Mengiste https://orcid.org/0000-0001-8865-994X

REFERENCES

Abebe, M. (2007). Nature and management of acid soils in Ethiopia. Addis Ababa: Ethiopian Institute for Agricultural Research.
Adugna, A. (2014). Analysis of in situ diversity and population structure in Ethiopian cultivated Sorghum bicolor (L.) landraces using
phenotypic traits and SSR markers. Springerplus, 3, 212. https://doi.org/10.1186/2193-1801-3-212
Agarwal, P. K., Agarwal, P., Jain, P., Jha, B., Reddy, M. K., & Sopory, S. K. (2008). Constitutive overexpression of a stress-inducible small GTP-binding protein PgRab7 from Pennisetum glaucum enhances abiotic stress tolerance in transgenic tobacco. Plant Cell Reports, 27, 105–115. https://doi.org/10.1007/s00299-007-0446-0
Alexander, D. H., & Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. Bmc Bioinformatics, 12, 246. https://doi.org/10.1186/1471-2105-12-246
Ambawat, S., Sharma, P., Yadav, N. R., & Yadav, R. C. (2013). MYB transcription factor genes as regulators for plant responses: An overview. Physiology and Molecular Biology of Plants, 19, 307–321. https://doi.org/10.1007/s12298-013-0179-1
An, J., Li, Q., Yang, J., Zhang, G., Zhao, Z., Wu, Y., & Wang, W. (2019). Wheat F-box protein TaFBA1 positively regulates plant drought tolerance but negatively regulates stomatal closure. Frontier Plant Science, 10, 1–20. https://doi.org/10.3389/fpls.2019.01242
Ayana, A., & Bekele, E. (2000). Geographical patterns of morphological variation in sorghum (Sorghum bicolor (L.) Moench) germplasm from Ethiopia and Eritrea: Qualitative characters. Euphytica, 205, 195–205. https://doi.org/10.1007/j.1601-5223.1998.101-1-00195.x
Ayana, A., Bryngelson, T., & Bekele, E. (2000). Genetic variation of Ethiopian and Eritrean sorghum (Sorghum bicolor (L.) Moench) germplasm assessed by random amplified polymorphic DNA (RAPD). Genetic Resources and Crop Evolution, 47, 471–482. https://doi.org/10.1023/a:100161-1623-0001-000249
Borrell, A. K., Mullet, J. E., George-jaeggi, B., & Van Oosterom, E. J. (2014). Drought adaptation of stay-green sorghum is associated with canopy development, leaf anatomy, root growth, and water uptake. Journal of Experimental Botany, 65, 6251–6263. https://doi.org/10.1093/jxb/eru232
Bort, J., Araus, J. L., Hazzam, H., Grando, S., & Salvatore, C. (1998). Relationships between early vigour, grain yield, leaf structure and stable isotope composition in field grown barley. Plant Physiology and Biochemistry, 36, 889–897.
Bouzroud, S., Gouiaa, S., Hu, N., Bernadac, A., Mila, I., Bendaou, N., … Zouine, M. (2018). Auxin Response Factors (ARFs) are potential mediators of auxin action in tomato response to biotic and abiotic stress (Solanum lycopersicum). Plos One, 13, 1–20. https://doi.org/10.1371/journal.pone.0193517
Briat, J.-F., Ravet, K., Arnaud, N., Ce lin Duc, J., Boucherez, O., Touraine, B., … Gaymard, F. (2010). New insights into ferritin and the environment. Toxins, 2, 105–115. https://doi.org/10.3390/toxins201-105
Brunet, J., & Larson-rabin, Z. (2012). The response of flowering time to global warming in a high-altitude plant: The impact of genetics and the environment. Botany, 90, 319–326. https://doi.org/10.1139/B2012-001
Califice, S., Baurain, D., Hanikenne, M., & Motte, Patrick (2012). A single ancient origin for prototypical serine/arginine-rich splicing factors. Plant Physiology, 158, 546–560. https://doi.org/10.1104/pp.111.189019
Chen, J., Morris, G. P., Chen, J., Chopra, R., Hayes, C., Morris, G., … Burrow, G. (2017). Genome-wide association study of heat tolerance of developing leaves during vegetative growth stages in a sorghum association panel. Plant Genome, 10, 1–15. https://doi.org/10.3835/plantgenome2016.09.0091
Cho, L., Yoon, J., & An, G. (2017). The control of flowering time by environmental factors. Plant Journal, 172, 708–719. https://doi.org/10.1111/tjp.13461
Cueva, H. E., & Prom, L. K. (2020). Evaluation of genetic diversity, agronomic traits, and anthracnose resistance in the NPGS Sudan Sorghum Core collection. Bmc Genomics [Electronic Resource], 21, 88. https://doi.org/10.1186/s12864-020-6489-0
Cueva, H. E., Rosa-Valentin, G., Hayes, C. M., Rooney, W. L., & Hoffmann, L. (2017). Genetic characterization of a core set of the USDA-NPGS Ethiopian sorghum germplasm collection: Implications for germplasm conservation, evaluation, and utilization in crop improvement. Bmc Genomics [Electronic Resource], 18, 108. https://doi.org/10.1186/s12864-016-3475-7
Cueva, H. E., Zhou, C., Tang, H., Khadke, P. P., Das, S., Lin, Y., … Paterson, A. H. (2016). The evolution of photoperiod-insensitive flowering in sorghum: A genomic model for panicoid grasses. Molecular Biology and Evolution, 33, 2417–2428. https://doi.org/10.1093/molbev/msw120
Dahlberg, J. A., Burke, J. J., & Rosenow, D. T. (2004). Development of a sorghum core collection: Refinement and evaluation of a subset from Sudan. Economic Botany, 58, 556–567. https://doi.org/10.1663/0013-0001(2004)058[0556:DOASCC]2.0.CO;2
Desmee, H., Jordan, D. R., & Godwin, I. D. (2016). Geographical patterns of phenotypic diversity in sorghum (Sorghum bicolor (L.) Moench) landraces from North Eastern Ethiopia. African Journal of Agricultural Research, 11, 3111–3122. https://doi.org/10.5897/AJAR2016.11121
Duncan, R., Bockholt, A., & Miller, F. R. (1981). Descriptive comparison of senescent and nonsenescent sorghum genotypes. Agronomy Journal, 73, 849–853. https://doi.org/10.2134/agronj1981.00021962007300505024x
Edwards, E. J., Osborne, C. P., Strömberg, C. A. E., & Smith, Stephen A. (2010). The origins of C4 grasslands: Integrating evolutionary and ecosystem science. Science, 328, 587–591. https://doi.org/10.1126/science.1177216
Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. Plos One, 6(5), e19379. https://doi.org/10.1371/journal.pone.0019379
Engels, J. M. M., & Hawkes, J. G. (1991). The Ethiopian gene centre and its genetic diversity. J. G. Hawkes, M. Worede, & J. M. M. Engels (Eds.). Cambridge: Cambridge University Press.
FAOSTAT. (2020). Statistical data base. FAO, Rome, Italy. Retrieved from http://www.fao.org/faostat/en/#data
Franco, J., Crossa, J., Taba, S., & Shands, H. (2005). A sampling strategy for conserving genetic diversity when forming core subset. Crop Science, 45, 1035–1044. https://doi.org/10.2135/cropsci2004.0292
Frankel, O. H. (1984). Genetic perspectives of germplasm conservation. E. Arber, W. K., Llimensee, K., Peacock, & W. J. and P. Straling, (Eds.). Cambridge: Cambridge University Press.
Fukuda, A., Chiba, K., Maeda, M., Nakamura, A., & Maeshima, M. (2004). Effect of salt and osmotic stresses on the expression of genes for the vacuolar H+-pyrophosphatase, H+-ATPase subunit A, and Na+/H+ antipporter from barley. Journal of Experimental Botany, 55, 585–594. https://doi.org/10.1093/jxb/erh070
Girma, G., Nida, H., Seyoum, A., Mekonen, M., Nega, A., Lule, D., … Mengiste, T. (2019). A large-scale genome-wide association analyses of Ethiopian sorghum landrace collection reveal loci associated with important traits. *Frontiers in Plant Science*, 10, 1–15. https://doi.org/10.3389/fpls.2019.00691

Gladman, N., Jiao, Y., Lee, Y. K., Zhang, L., Chopra, R., Regulski, M., … Xin, Z. (2019). Fertility of pediculate spikelets in sorghum is controlled by a jasmonic acid regulatory module. *International Journal of Molecular Sciences*, 20, 1–15. https://doi.org/10.3390/ijms20194951.

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., & Buckler, E. S. (2014). TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. *PloS One*, 9, e90346. https://doi.org/10.1371/journal.pone.0090346

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. R., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., & Rohksar, D. S. (2012). Phytozone: A comparative platform for green plant genomics. *Nucleic Acids Research*, 40, D1178–D1186. https://doi.org/10.1093/nar/gkr944

Grenier, C., & Hamon, P. (2001). Core collection of sorghum: I. Stratification based on eco-geographical data. *Crop Science*, 41, 379–380.

Hu, Z., Olatoye, M. O., Marla, S., & Morris, G. P. (2019). An integrated genotyping-by-sequencing polymorphism map for over 10,000 sorghum genotypes. *Plant Genome*, 12, 1–15. https://doi.org/10.3835/plantgenome2018.06.0044

Huang, M., Liu, X., Zhou, Y., Summers, R. M., & Zhang, Z. (2018). B.LINK: a package for the next level of genome-wide association studies with both individuals and markers Meng Huang. *Gigascience*, 8, 1–12. https://doi.org/10.1093/gigascience/giy154

Hummer, K. E., & Hancock, J. F. (2015). Vavilovian centers of plant diversity: Implications and impacts. *HortScience*, 50, 780–783. https://doi.org/10.21273/HORTSCI.50.6.780

IBC. (2007). Ethiopia: Second country report on the state of PGRFA to FAO. Addis Ababa: Ethiopia.

IBPGR and ICRISAT. (1993). Descriptors for sorghum (*Sorghum bicolor* (L) Moench). International Board for Plant Genetic Resources, Rome, Italy; International Crops Research Institute for the Semi-Arid Tropics, Patancheru, India.

Imran, M., Tang, K., & Liu, J. (2016). Comparative Genome-wide analysis of the malate dehydrogenase gene families in cotton. *PloS One*, 11, 1–22. https://doi.org/10.1371/journal.pone.0166341

Janiak, A., Kwasniewski, M., Sowa, M., & Gajek, K. (2018). No time to waste: Transcriptome study reveals that drought tolerance in barley may be attributed to stressed-like expression patterns that exist before the occurrence of stress. *Frontiers in Plant Science*, 8. https://doi.org/10.3389/fpls.2017.02212

Kottek, M., Jurgen, G., Christoph, B., Bruno, R., & Franz, R. (2006). World map of the Köppen-Geiger climate classification updated. *Meteorologische Zeitschrift*, 15, 259–263. https://doi.org/10.1127/0941-2948/2006/0130

Kumar, K., Rao, K. P., Biswas, D. K., & Sinha, A. K. (2011). Rice WNK1 is regulated by abiotic stress and involved in internal circadian rhythm. *Plant Signaling & Behavior*, 6, 316–320. https://doi.org/10.4161/psb.6.3.31063

Kumari, P., Pahuja, S. K., Arya, S., & Patil, J. V. (2016). Sorghum. In *Broadening the genetic base of grain cereals* (pp. 163–203). Springer, New Delhi.

Leff, B., Ramankutty, N., & Foley, J. A. (2004). Geographic distribution of major crops across the world. *Global Biogeochemical Cycles*, 18, 1–27. https://doi.org/10.1029/2003GB002108

Li, F., Wen, W., Liu, J., Zhang, Y., Cao, S., He, Z., … Jin, H. (2019). Genetic architecture of grain yield in bread wheat based on genome-wide association studies. *Bmc Plant Biology [Electronic Resource]*, 19, 168. https://doi.org/10.1186/s12870-019-1781-3

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

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

Lopez-Hernandez, F., & Cortes, A. J. (2020). Last-generation genome: Environment associations reveal the genetic basis of heat tolerance in common bean (*Phaseolus vulgaris L*). *Frontiers in Genetics*, 10, 1–22. https://doi.org/10.3389/fgene.2019.00954

Mace, E. S., Buhariwalla, K. K., Buhariwalla, H. K., & Crouch, J. H. (2003). A high-throughput DNA extraction protocol for tropical molecular breeding programs. *Plant Molecular Biology Reporter*, 21, 459–460. https://doi.org/10.1007/BF02772596

Maina, F., Bouchet, S., Marla, S. R., Hu, Z., Wang, J., Mamadou, A., … Morris, G. P. (2018). Population genomics of sorghum (*Sorghum bicolor*) across diverse agroclimatic zones of Niger. *Genome*, 61, 223–232. https://doi.org/10.1139/gen-2017-0131

McCormick, R. F., Truong, S. K., Reedasayam, A., Jenkins, J., Shu, S., Sims, D., … Mullet, J. E. (2018). The sorghum bicolor reference genome: Improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *Plant Journal*, 93, 338–354. https://doi.org/10.1111/tpj.13781

Mengistu, G., Shimelis, H., Laing, M., & Lule, D. (2019). Assessment of farmers’ perceptions of production constraints, and their trait preferences of sorghum in western Ethiopia: Implications for anthracnose resistance breeding. *Acta Agriculturae Scandinavica. Section B. Soil and Plant Science*, 69, 241–249. https://doi.org/10.1080/09064710.2018.1541190

Morris, G. P., Ramu, P., Deshpande, S. P., Hash, C. T., Shah, T., Upadhyaya, H. D., … Kresovich, S. (2013). Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proceedings of the National Academy of Sciences*, 110, 453–458. https://doi.org/10.1073/pnas.1215985110

Mundia, C. W., Secchi, S., Akamani, K., & Wang, G. (2019). A regional comparison of factors affecting global sorghum production: The case of North America, Asia and Africa’s Sahel. *Sustainability*, 11, 1–18. https://doi.org/10.3390/su11072135

Mutava, R. N., Prasad, P. V. V., Tuinstra, M. R., Kofoid, K. D., & Yu, J. (2011). Field crops research characterization of sorghum genotypes for traits related to drought tolerance. *Field Crops Research*, 123, 10–18. https://doi.org/10.1016/j.fcr.2011.04.006

Nida, H., Girma, G., Mekonen, M., Lee, S., Seyoum, A., Dessalegn, K., … Mengiste, T. (2019). Identification of sorghum grain mold resistance loci through genome wide association mapping. *Journal of Cereal Science*, 85, 295–304. https://doi.org/10.1016/j.jcs.2018.12.016

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20, 289–290. https://doi.org/10.1093/bioinformatics/20.1.289

Polo, M. J., Liu, Y., Lee, Y. K., Zhang, L., Chopra, R., Regulski, M. A., & Buckler, E. S. (2012). Fast and accurate genome-wide association analysis pipeline. *Gigascience*, 259–263. https://doi.org/10.1093/gigascience/gts006
Plant Genetic Resources Center. (1995). Ethiopia: Country report to the FAO international technical conference on plant genetic resources (Leipzig, 1996). Addis Ababa: Ethiopia.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics, 81, 559–575. https://doi.org/10.1086/519795

R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation of Statistical Computing. Retrieved from http://www.R-project.org/

Rhodes, D. H., Hoffmann, L., Rooney, W. L., Herald, T. J., Bean, S., Boyles, R., … Kresovich, S. (2017). Genetic architecture of kernel composition in global sorghum germplasm. Bmc Genomics [Electronic Resource], 18, 15. https://doi.org/10.1186/s12864-016-3403-x

Shannon, C. E., & Weaver, W. (1949). The Mathematical Theory of Communication. Urbana, IL: The University of Illinois Press.

Silva, M. D. A., Jifon, J. L., Da Silva, J. A. G., & Sharma, V. (2007). Use of physiological parameters as fast tools to screen for drought tolerance in sugarcane. Brazilian Journal of Plant Physiology, 19, 193–201. https://doi.org/10.1590/S1677-04202007000300003

Singh, S. P. (1985). Sources of cold tolerance in grain sorghum. Canadian Journal of Plant Science, 65, 251–257. https://doi.org/10.4141/cjps85-037

Stephens, J. C., Miller, F. R., & Rosenow, D. T. (1967). Conversion of alien sorghums to early combine genotypes. Crop Science, 7, 396–396. https://doi.org/10.2135/cropsci1967.0011183X000700040036x

Studnicki, M., & Debski, K. (2012). ccChooser: Developing a core collections. Retrieved from http://cran.r-project.org/package=ccChooser

Tari, I., Laskay, G., Takacs, Z., & Poor, P. (2013). Response of sorghum to abiotic stresses: A review. Crop Science, 199, 264–274. https://doi.org/10.1111/jac.12017

Torre, A. R. D. La, Wilhite, B., & Neale, D. B. (2019). Environmental genome-wide association reveals climate adaptation is shaped by subtle to moderate allele frequency shifts in loblolly pine. Genome Biology and Evolution, 11, 2976–2989. https://doi.org/10.1093/gbe/evz220

Turner, S. D. (2014). qman: an R package for visualizing GWAS results using Q-Q and manhattan plots. Journal of Open Source Software, 3, 731. https://doi.org/10.21055/joss.00731

Upadhyaya, H. D., Sharma, S., Dwivedi, S., & Singh, S. (2014). Sorghum genetic resources: Conservation and diversity assessment for enhanced utilization in sorghum improvement. U.H. and K. C. Wang Y-H, (Eds.), Boca Raton, FL: CRC Press (Taylor & Francis).

Wang, M., Yan, J., Zhao, J., Song, W., Zhang, X., Xiao, Y., & Zheng, Y. (2012). Genome-wide association study (GWAS) of resistance to head smut in maize. Plant Science, 196, 125–131. https://doi.org/10.1016/j.plantsci.2012.08.004

Wang, Y., Liu, K., Liao, H., Zhuang, C., Ma, H., & Yan, X. (2008). The plant WNK gene family and regulation of flowering time in Arabidopsis. Plant Biology, 10, 548–562. https://doi.org/10.1111/j.1438-8677.2008.00072.x

Weerasooriya, D. K., Maulana, F. R., Bandara, A. Y., Tirfessa, A., Mengistu, G., Nohou, K., & Tesso, T. T. (2016). Genetic diversity and population structure among sorghum (Sorghum bicolor L.) germplasm collections from western Ethiopia. African Journal of Biotechnology, 15, 1147–1158. Retrieved from https://doi.org/10.5897/AJB2015.14604

Wickham, H. (2016). Ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag.

Wortmann, C. S., Mamo, M., Mburu, C., Letayo, E., Abebe, G., Kayuki, K. C., … Ndacyayisenga, T. (2009). Lincoln atlas of sorghum (Sorghum bicolor (L.) Moench): Production in eastern and southern Africa. Lincoln, NE: University of Nebraska.

Yan, W., Rutger, J. N., Bryant, R. J., Bockelman, H. E., Fjellstrom, R. G., Chen, M., … Mcclung, A. M. (2007). Development and evaluation of a core subset of the USDA rice germplasm collection. Crop Science, 47, 869–878. https://doi.org/10.2135/cropsci2006.07.0444

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M., & Buckler, E. S. (2010). Mixed linear model approach adapted for genome-wide association studies. Nature Genetics, 42, 355–360. https://doi.org/10.1038/ng.546

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Girma G, Nida H, Tirfessa A, et al. A comprehensive phenotypic and genomic characterization of Ethiopian sorghum germplasm defines core collection and reveals rich genetic potential in adaptive traits. Plant Genome. 2020;13:e20055. https://doi.org/10.1002/tpg.220055