Genet Resour Crop Evol (2023) 70:1439–1454
https://doi.org/10.1007/s10722-022-01513-4

RESEARCH ARTICLE

Genomics-based assembly of a *sorghum bicolor* (L.) moench core collection in the Uganda national genebank as a genetic resource for sustainable sorghum breeding

R. Mufumbo · S. Chakrabarty · M. Nyine · S. M. Windpassinger · J. W. Mulumba · Y. Baguma · L. T. Odong · M. Frisch · R. J. Snowdon

Received: 3 June 2022 / Accepted: 3 December 2022 / Published online: 21 December 2022

Abstract The Uganda National GeneBank is a key reservoir of genetic diversity for sorghum (*Sorghum bicolor* (L.) Moench), with over 3333 accessions which are predominantly landraces (96.48%), but also includes the weedy accessions (0.63%), breeding lines (2.5%) and released varieties (0.39%). This genetic resource from the primary center of sorghum diversity and domestication is important for broadening the genetic diversity of elite cultivars through breeding. However, due to the large size of the collection, we aimed to select a core set that captures the maximum genetic and phenotypic diversity, in order to facilitate detailed genetic and phenotypic evaluation at a reduced cost. To achieve this, we genotyped the entire collection in 2020 using Diversity Array Technology sequencing (DArTseq). A total of 27,560 SNPs were used to select a core collection of 310 accessions using the GenoCore software. A comparison of core set and the whole collection based on the polymorphism information content, observed heterozygosity, expected heterozygosity and minor allele frequency showed no significant difference between the two sets, indicating that the core collection adequately captures the genetic diversity and allelic richness present in the whole collection. The core collection captures all the five major sorghum races and the 10 intermediate hybrids. The most strongly represented race is guinea (24.5%), while caudatum-bicolor is least frequent (0.69%). Landraces account for 92.2% of the core collection, whereas breeder's

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10722-022-01513-4.

R. Mufumbo · S. Chakrabarty · S. M. Windpassinger · R. J. Snowdon (✉)
Department of Plant Breeding, Justus Liebig University Giessen, 35392 Giessen, Germany
e-mail: rod.snowdon@agrar.uni-giessen.de

R. Mufumbo · J. W. Mulumba
Uganda National Gene Bank, Biodiversity and Biotechnology Program, National Agricultural Research Laboratories-Kawanda, P.O. Box 7065, Kampala, Uganda

M. Nyine
Department of Plant Pathology, Kansas State University, Manhattan, KS 66506, USA

Y. Baguma
National Agricultural Research Organization (NARO) Secretariat Entebbe, P.O. Box 295 Berkeley Rd, Entebbe, Uganda

L. T. Odong
Department of Agricultural Production, College of Agricultural and Environmental Sciences, Makerere University, P. O. Box 7062, Kampala, Uganda

M. Frisch
Institute of Agronomy and Plant Breeding II, Justus Liebig University Giessen, 35392 Giessen, Germany

lines, weedy accessions and released varieties contribute 2.2%, 3.5% and 1.9%, respectively.

Keywords Core collection · Genetic diversity · GenoCore · Germplasm conservation · Sorghum

Introduction

Sorghum bicolor (L.) Moench (hereinafter referred to as sorghum) is the fifth most produced cereal globally and the second most widely grown cereal in Africa behind maize. In 2020, an estimated total of 28.14 M t of sorghum grain was produced across all of Africa on 22.46 M ha (https://www.fao.org/faostat/en/#data/). In Uganda, sorghum is the third most important cereal crop after maize and rice, with an estimated production area in 2020 of 305,721 ha and a total production volume of 251,634 t. Most sorghum in Uganda is produced by smallholder farmers. It is a staple food in the cool Kigezi highlands and the semi-arid regions of Eastern and Northern Uganda.

Uganda is located in the primary center of sorghum genetic diversity and domestication, an area that extends from Ethiopia to Sudan and the surrounding countries of East Africa (Doggett 1965; Mukuru 1993). It is believed that cultivated races of sorghum, were domesticated in East Africa around 1000BC, possibly along the Nile where the greatest diversity of this species is still found (Damon 1962; Kimber 2000). Uganda is one of the three countries globally in which all the five basic sorghum races and ten intermediate races are endemic (Reddy et al. 2002), making it an important source of genetic diversity for sorghum breeding. The high genetic diversity of sorghum found in Uganda provides a potential source of novel alleles for improving pathogen and pest resistances, tolerance to abiotic stresses such as heat, cold and drought, yield and other complex agronomic traits such as end-use quality of food, feed and industrial products.

Much of Uganda's sorghum genetic diversity is conserved in the Uganda National GeneBank in Entebbe (http://www.pgrc.go.ug). The sorghum collection in this gene bank was established to safeguard valuable germplasm from genetic erosion as a result of habitat loss, climate change and the increasing adoption of modern varieties by farmers. The conserved sorghum germplasm captures thousands of years of evolutionary history, making Uganda a critical reservior for mining novel alleles that are urgently needed for sorghum improvement. The accessions in the collection represent a broad ecological adaptation with an ability to grow in diverse climates, including the semi-arid areas in Karamoja which receive very little annual precipitation, the high rainfall and humid areas of Busoga in the Lake Victoria crescent, and the colder areas of the Kigezi highlands.

The Uganda National GeneBank currently conserves 3333 sorghum accessions representing all sorghum agro-ecological zones and ´ethno-cultural diversity in Uganda. Managing such a large collection is costly, time-consuming and labor-intensive. In addition, in-depth evaluation at molecular and phenotypic level is equally costly, thus limiting its utility in sorghum breeding programs. To date, the germplasm has not been systematically characterized and evaluated for complex quantitative traits.

The value of a germplasm collection is determined not by its size but by how well the genetic and phenotypic diversity are characterized, documented and made available to breeders and other users. The inherent challenge with many germplasm collections in genebanks worldwide is redundancy, which is usually caused by use of different synonyms during germplasm collection expeditions. The most efficient approach to manage and use large genebank collections is to identify a core set that captures the maximum genetic diversity available in the genebank (Frankel 1984; Brown 1989a; van Hintum et al. 2000). In many genebanks, core collections comprising 10% of the entire collection represent a more manageable number of accessions which is easier to maintain and utilize (Frankel 1984; Frankel and Brown 1984).

In large genebanks with many accessions, the core collection may still be relatively large, thus miniature ("mini") core collections comprising 1% of the core collection have been considered as an alternative (Upadhyaya and Ortiz 2001). These subsets in genebanks that are genetically diverse serve as panels for extensive evaluation of important agronomic, disease and pest resistance, and abiotic tolerance traits under replicated multi-locational trials. This allows for efficient generation of information that serves as a guide for more efficient use of the entire collection in the crop breeding programs (Brown 1989b). Among other applications, core collections are particularly

relevant for gene discovery and allele mining through genotyping by sequencing (GBS) or whole genome re-sequencing (Balfourier et al. 2007; Richards et al. 2009). Genetic markers and phenotype data generated from core collections facilitate genetic association mapping studies (Le Cunff et al. 2008; El Bakkali et al. 2013), and the identification of interesting parents for generating biparental populations for linkage mapping studies (Barnaud et al. 2006; Cubry et al. 2013). Several methods for assembling core collections are available including MSTRAT (Gouesnard et al. 2001), GenoCore (Jeong et al. 2017), Core Hunter (Thachuk et al. 2009; Beukelaer et al. 2012), principal component scoring (Noirot et al. 1996) and the distance-based methods such as MLST (Perrier et al. 2003) and Power Core (Kim et al. 2007). The fundamental principle behind the methods is the ability to maximize allelic diversity/richness in a reduced sample size. The choice of the method depends on the purpose of the study (for example capturing maximum variation vs. optimizing the chance of finding new alleles), computational speed and the requirement for a priori information (e.g., preselected markers, defined subgroups and/or sample size) (Odong et al. 2013). Distance based methods mainly aim at maximizing allelic diversity at the genome level, which is suitable in breeding (Leroy et al. 2014), whereas methods that capture the highest number of alleles including the rare alleles are more suitable for germplasm conservation (Schoen and Brown 1993).

The main aim of this study was to assemble and evaluate a core collection that captures the maximum allelic richness among the 3333 sorghum accessions in the Uganda National GeneBank, so that breeders and other users can extract genotypes suitable for crop improvement, genetic analyses of interesting traits and other purposes.

Materials and methods

Germplasm

A total of 3333 sorghum accessions from the Uganda National GeneBank in Entebbe were used in this study (Supplementary Table S1). The collection consists of landraces (96.48%), improved varieties (0.39%), breeding lines (2.5%) and weedy accessions (0.63%) (http://www.pgrc.go.ug).

DNA extraction and genotyping

A single representative seed per accession was used for DNA extraction. The seeds were ground into fine powder which was shipped to Diversity Arrays Technology (DArT) Pty Ltd, Canberra Australia (http://www.diversityarrays.com/dart-mapsequences) for sequence-based, genome-wide DArTseq genotyping (Diversity Arrays Technology Pty, Canberra, Australia). Library preparation, sequencing and read processing were performed by the service provider following proprietary protocols for DArTseq. Sequence reads were aligned to the sorghum reference genome assembly BTx623 version 3 available at https://phytozome-next.jgi.doe.gov/info/Sbicolor_v3_1_1 (McCormick et al. 2017). The alignment thresholds were E-value = 5e– 5 and minimum percent identity = 70%. Basic statistics including call rate, minor allele frequency, major allele frequency and heterozygosity were estimated.

Core collection assembly

GenoCore software (Jeong et al. 2017) was used to select entries for the Uganda national sorghum core collection based on 27,560 SNP markers retained after filtering out SNP sites with more than 20% missing data. GenoCore was chosen because of it speed and consistence when handling large datasets. We set the parameters coverage (– cv) to 100% and delta (– d) to 0.01% to ensure that the accessions selected by GenoCore reflected the diversity in the whole collection. Therefore, the size of the final core collection was determined by the level of genetic diversity present in the whole genebank collection rather than being set a priori.

Genetic diversity analysis

A custom Perl script was used to convert the DArTSeq single row marker data into a wide nucleotide base format with accessions as rows and SNPs as columns. The raw SNP dataset was filtered using the R package snpReady v0.9.6 (Granto et al. 2018) with parameters call.rate = 0.9, maf = 0.01 and sweep.sample = 0.5. Nine samples with more than 50% missing data were excluded and a total of 8251 unique SNPs out of the 39,933 genotyped SNP sites were retained. The snpReady function ‘popgen’ was used

to recalculate the genetic diversity statistics including minor allele frequency (MAF), expected heterozygosity (He), observed heterozygosity (Ho), Nei's genetic diversity (GD) and polymorphism information content (PIC). Wright's fixation index (F_{st}) was calculated according to Granato and Fritsche-Neto (2017). Additional genetic diversity parameters such as Tajima's D were calculated using TASSEL v5.0 (Bradbury et al. 2007). The analysis was performed on the whole population, core collection and subpopulations stratified by biological status, geographical regions and districts. Hierarchical clustering of the core collection was done using hclust function of the R package 'ape' based on the Nei's genetic distances calculated using the function nei.dist provided in R package poppr (Kamvar et al. 2014).

Population structure

To understand the population structure and the proportion of ancestry admixture within the sorghum collection maintained by the Uganda National GeneBank, we used both the principal component analysis (PCA) and ancestry analysis implemented in the program ADMIXTURE v1.3.0 (Alexander et al. 2015), respectively. A total of 7091 SNPs retained after filtering out SNPs with call.rate < 0.9, maf < 0.05 and maximum missing data less than 10% was used to calculate the principal components using the 'prcomp' function in R. The missing genotypes were imputed using beagle v5.0 (Browning and Browning, 2013) before PCA was done. The first two principal components were plotted using ggplot2 (Wickham 2016).

For ancestry analysis, a custom Perl script was used to convert the filtered SNPs into a hapmap format whereas the TASSEL pipeline (Glaubitz et al. 2014) was used to convert the filtered SNP hapmap file to plink format. The final input files for ADMIXTURE analysis were prepared using the plink software (http://pngu.mgh.harvard.edu/purcell/plink; Purcell et al. 2007). We tested 12 K values to determine the optimal number of clusters within the population. A plot of K against cross validation errors and the knowledge of geographical and biological status stratification were used to determine the best K. Stacked bar plots were generated to show the level of admixture between accessions after sorting the Q values and the pairwise F_{st} values were recorded.

Validation of the core collection

The degree to which the core collection represents the entire germplasm collection was validated by comparing the diversity parameters such as MAF, PIC, Ho and He for the whole collection and the core collection, respectively. PCA was also conducted to confirm whether the core collection represented the genetic diversity of the whole sorghum collection maintained in the national genebank.

Phenotype variation in the core collection

The core collection was characterized at Puerto Vallarta 20° 3912.2652 N, 105° 1331.1952 W in Mexico on the Pacific Coast, in un-replicated nursery micro plots, following local good agronomical practices. The phenotypic diversity in sorghum is associated with adoption and use of accessions in different cultural and agro-ecological zones. Therefore, a minimal descriptor of phenotypic diversity captured in the core collection was evaluated using traits such glume colour, grain colour, race, percentage of grain covered by the glume and days to 50% flowering. Racial classification was based on morphological criteria (spikelet structure and panicle shape) (Harlan and De Wet. 1972).

Results

Core collection composition

A total of 310 entries were assigned to the core collection, representing approximately 10% of the full sorghum collection of the Uganda National GeneBank. The distribution of accessions from the core relative to the whole collection on the first two principal components shows that the core collection is a good representative of the sorghum genetic diversity in the Uganda National GeneBank because it captures the regional gene pools and biological status (Fig. 1; Table 1). The first two principal components explained 23% of the genetic variance found in the whole collection. Accessions from Northern and Northwestern regions clustered together resulting in three distinct groups corresponding to germplasm collected in Northern/Northwestern, Eastern and Southwestern regions of Uganda, respectively.

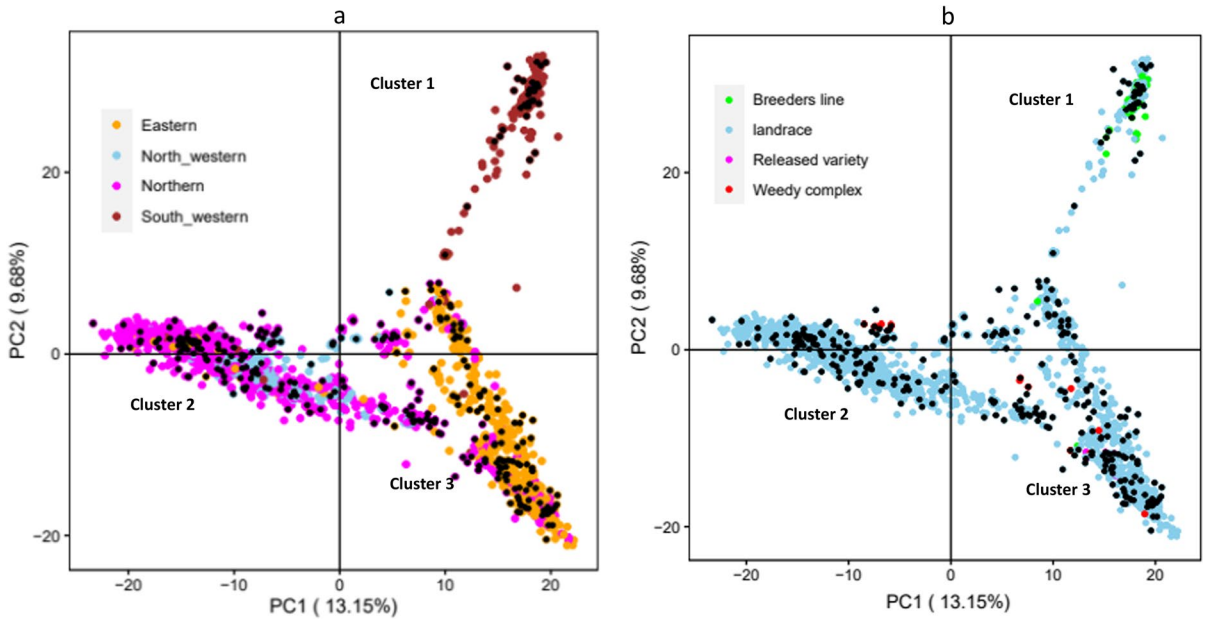

Fig. 1 Distribution of sorghum accessions maintained in the Uganda National GeneBank on the first two principal components. Where **a** shows the clustering based on geographical regions whereas **b** show the clustering based on the biological status. The black dots represent the 310 accessions in the core collection

Northern Uganda, particularly Gulu and Omoro districts, contributed most to the core collection, whereas the cold Kigezi highlands in the Southwestern region contributed the least number of accessions to the core collection (Table 1). In general, the core collection was dominated by landraces with about 92.3% of the total accessions, whereas the remaining accessions represented the weedy accessions, breeding lines and released varieties.

Genetic diversity and representativeness of the core collection

There was no significant difference in MAF between the core collection and the whole population ($P = 0.2333$) based on the t-test at 5% confidence level (Table 2). The core collection showed a two-fold difference in Ho (0.062 ± 0.0006) when compared to the whole population (0.033 ± 0.0006) and other subpopulations, with the exception of the weedy accessions (Table 3), suggesting that a high level of genetic variability was captured. The level of PIC for the core collection did not vary much from that of the whole population, averaging at 0.2. The F_{st} values for the whole collection and the core collection were 0.000 and 0.019, respectively, confirming a high level of random mating between subpopulations due to a lack of genetic isolation. The breeder's lines and the Southwestern subpopulation had the highest F_{st} values (0.438 and 0.365, respectively), suggesting that they are more differentiated from other subpopulations. These results are consistent with the population structure revealed by PCA (Fig. 1), which indicates that majority of the breeder's lines were collected from the Southwestern subpopulation. However, the level of genetic differentiation is not strong enough to prevent crossbreeding with other populations. Interestingly, a negative Tajima's D (– 0.799) was observed only in the breeder's lines, suggesting the presence of selection-sweeps through breeding, whereas other subpopulations had positive Tajima's D values, indicating varying levels of balanced selection.

Ancestry analysis of the core collection

Based on the changes in the cross-validation error, there was no clear-cut number of clusters in the whole collection due to the high level of shared ancestry (Fig. 2a). However, K = 4 seemed

Table 1 Composition of the core collection in terms of origin and biological status of the sorghum germplasm. Bukedi under Eastern region includes Tororo, Busia, Budaka, Kibuku, Butaleja and Pallisa districts

District /Origin	Accessions in Whole collection	Proportion in the whole collection (%)	Accessions in core collection	Final contribution to the core Collection (%)
Eastern				
Bukedi*	228	17.9	41	13.2
Soroti	56	10.7	6	1.9
Katakwi	67	14.9	10	3.2
Kumi	86	9.3	8	2.5
Ngora	39	10.2	4	1.3
Iganga	105	11.4	12	3.8
	581	17.42	81	26.2
Northern				
Kole	64	14.1	9	2.9
Nwoya	95	12.6	12	3.8
Pader	186	5.9	11	3.5
Kitgum	83	9.6	8	2.5
Otuke	57	5.2	3	0.9
Gulu	332	10.2	34	11
omoro	305	10.49	32	10.3
Amuru	39	2.5	1	0.3
Alebtong	100	10	10	3.2
Agago	34	8.8	3	0.97
Lira	221	5.4	12	3.8
	1516	45.45	135	43.6
Southwestern higlands				
Kabale	118	6.7	8	2.5
Rukiga	25	12	3	0.9
Rubanda	85	7	6	1.9
	228	25.7	17	5.5
North-western (West Nile)				
Zombo	151	1.9	3	0.9
Yumbe	181	7.1	13	4.2
Arua	169	4.1	7	2.2
	501	**13.1**	**23**	**7.4**
Research Institutes				
Makerere University	103	3.0	15	4.8
Nabuinzardi (Moroto)	63	1.8	8	2.5
NaSARRI (Serere)	10	0.002	5	1.6
Abizardi (Arua)	17	0.5	5	1.6
Kazardi (Kabale)	151	4.5	19	6.1
	344	10.31	52	16.6
Biological status	Accessions in whole collection	Proportion in whole collection (%)	Accessions in core collection	Contribution to core collection (%)
Landraces	3216	96.5	285	92.2
Weedy accessions	21	0.63	11	3.5
Breeding lines	83	2.5	7	2.2

Table 1 (continued)

Biological status	Accessions in whole collection	Proportion in whole collection (%)	Accessions in core collection	Contribution to core collection (%)
Released varieties	13	0.4	6	1.9
Total	3333		310	

Table 2 Genetic diversity Indices for the sorghum collection in the Uganda National GeneBank

Group	No. of accs	MAF	He	Ho	PIC	F_{st}	Tajima's D
Whole collection	3333	0.184 ± 0.0016	0.260 ± 0.0017	0.033 ± 0.0006	0.213 ± 0.0013	0.000	3.505
Core collection	310	0.181 ± 0.0015	0.255 ± 0.0017	0.062 ± 0.0006	0.209 ± 0.0013	0.019	2.443
Breeding line	83	0.100 ± 0.0017	0.146 ± 0.0021	0.034 ± 0.0014	0.123 ± 0.0016	0.438	– 0.799
Landrace	3216	0.182 ± 0.0016	0.258 ± 0.0017	0.034 ± 0.0006	0.212 ± 0.0013	0.008	3.438
Weedy accessions	21	0.237 ± 0.0017	0.322 ± 0.0017	0.108 ± 0.0013	0.258 ± 0.0012	– 0.238	0.940
Eastern	676	0.186 ± 0.0017	0.258 ± 0.0019	0.040 ± 0.0007	0.210 ± 0.0014	0.008	2.627
Northwestern	526	0.162 ± 0.0016	0.233 ± 0.0019	0.030 ± 0.0009	0.192 ± 0.0014	0.104	1.838
Northern	1731	0.165 ± 0.0015	0.237 ± 0.0018	0.041 ± 0.0007	0.196 ± 0.0013	0.088	2.619
Southwestern	385	0.111 ± 0.0016	0.165 ± 0.0019	0.036 ± 0.0012	0.140 ± 0.0014	0.365	0.345

MAF Major allele frequency, *He* Expected heterozygosity, *Ho* Observed heterozygosity, *PIC* Polymorphic Information Content, F_{ST} Inbreeding coefficient

Table 3 Genetic differentiation (F_{st}) between subpopulations at K = 4 with regions contributing the greatest number of accessions to the group indicated in the parentheses

	Pop1 (N)	Pop2 (NW)	Pop3 (SW)
Pop1 (Northern)			
Pop2 (Northwestern)	0.337		
Pop3 (Southwestern)	0.578	0.637	
Pop4 (Eastern)	0.44	0.505	0.609

CV error (K = 4): 0.47877, Loglikelihood: -14,030,780.139746

reasonable, as it separated the accessions from Southwestern, Northern and Northwestern, Eastern and the overlap between the Eastern and Northern subpopulations (Fig. 2b), which was in agreement with the results of PCA (Fig. 1). The level of admixture between the Eastern and Northern subpopulation was high, suggesting a continuous gene flow between these two regions. The pairwise F_{st} values between estimated populations at K = 4 varied between from 0.337 to 0.637, indicating that the level of genetic differentiation is low for reproductive isolation to occur due to an ongoing crossbreeding between subpopulations (Table 3). The Southwestern subpopulation showed the highest level of genetic differentiation from other populations, although the F_{st} value of 0.637 between south western and north western subpopulations was high enough to cause reproductive isolation.

Similarly, hierarchical clustering of accessions from the core collection based on Nei's genetic distances also revealed four main clusters corresponding to geographical regions of origin (Fig. 3, Supplementary Table S2). In the Northern subpopulation, fourteen landraces formed a unique cluster (N*) and these were collected from Omoro district. The accessions show very low admixture with the Eastern and Southwestern subpopulation, indicating restricted geneflow between these landraces and other germplasm except those from the Northwestern region. The Southwestern subpopulation also showed a very low admixture with other subpopulations. The breeding lines selected from the Southwestern population served as a genetic bridge between the Southwestern gene pool and the Northern and Eastern subpopulations through crossbreeding. However, the majority of the accessions from the Southwestern region are genetically distinct and appear to have no shared ancestry with the Northwestern subpopulation (Fig. 3B).

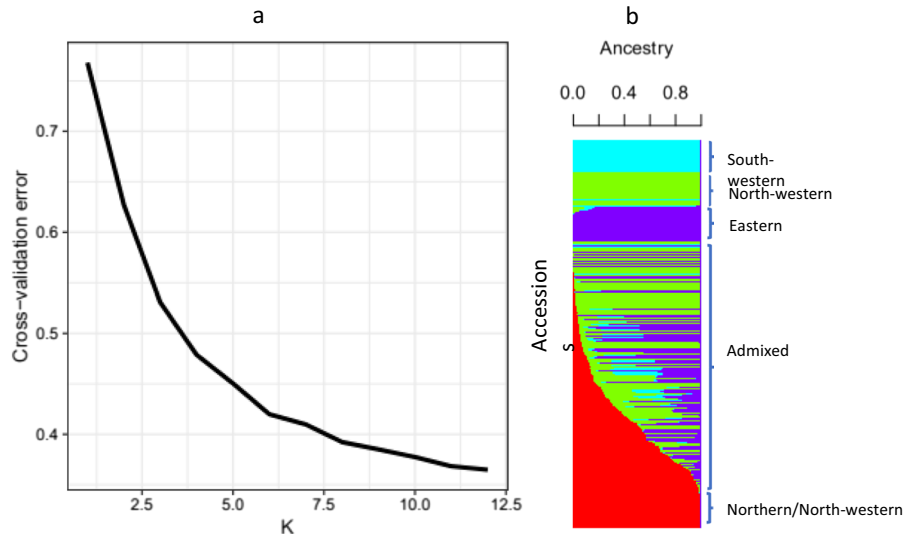

Fig. 2 Population structure of the sorghum germplasm collection in Uganda National GeneBank showing a high level of ancestry admixture. **a** shows the relationship between the K and cross-validation errors whereas **b** is a stacked bar plot showing ancestry admixture among all accessions

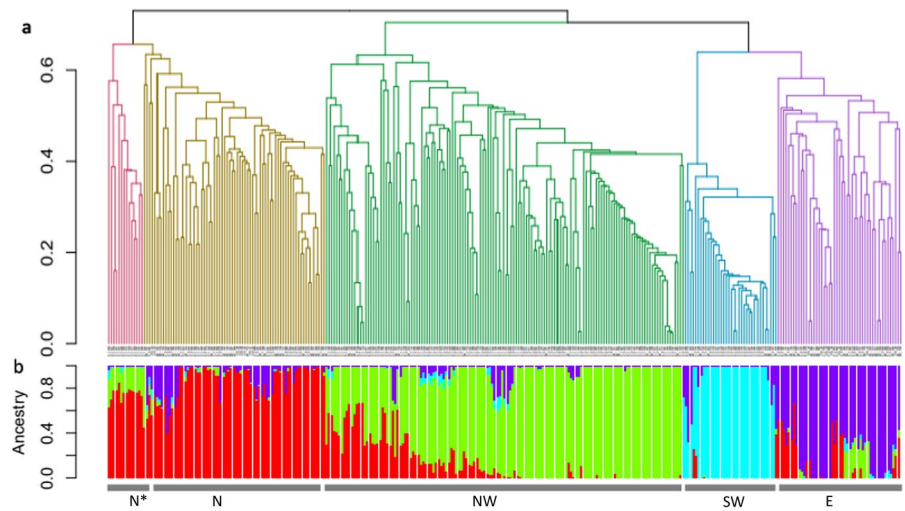

Fig. 3 Hierarchical clustering and the level of ancestry admixture of accessions in the core collection of *Sorghum bicolor* germplasm in Uganda. **a** Dendrogram showing clustering of accessions based on Nei's genetic distance. **b** Stack barplot showing ancestry admixture in the accessions. Geographical origin of accessions are indicated as follows: *N* Northern; *NW* Northwestern; *SW* Southwestern; *E* Eastern, *N** Omoro District in Northern Uganda

Agro-morphological diversity of the core collection

Racial diversity

The sorghum races captured by the core collection reflected the full *S. bicolor* racial spectrum. All five major sorghum races and their 10 intermediate hybrids are represented in the core collection (Fig. 3, Table 4, Supplementary Table S3). The most represented race was guinea (24.5%) whereas the least represented race was caudatum (0.69%), (Table 4). The northern region had the highest sorghum racial diversity, with all five major races and 9 intermediate races, whereas the cold Kigezi highland region had the least, with only 3 main races and 6 intermediate races (Supplementary Table S3). Interestingly, all the 18 weedy accessions belonged to the guinea race and were from northern Uganda.

Grain color diversity

There was a high seed color diversity in the core collection (red, brown, buff, yellow, orange, purple and white) (Table 4). Of the seven seed color classes, red was dominant (40.2%) whereas orange (0.69%), purple (1.4%) and (yellow 1.7%) had the lowest frequencies. The Northern region had the highest seed color diversity, with all the seven seed

Table 4 Variability of the quantitative and qualitative panicle and grain traits of the core collection. Information is based on 286 core collection accessions, excluding 24 accessions which did not germinate in this trial

Traits	Variants	%
Glume colour	Black	23.8
	Brown	28.7
	Buff	21.7
	Purple	10.5
	Red	15.4
Grain colour	Red	40.2
	Brown	25.5
	White	16.4
	Buff	13.6
	Purple	1.4
	Yellow	1.7
	orange	1
Glume covering	25%	65.7
	50%	23.1
	75	9.1
	Grain fully covered	1.1
	Glume longer than grain	1.1
Race	Bicolor	1.4
	Guinea	24.5
	Caudatum	4.1
	Kafir	4.5
	Durra	2.4
	Guinea-Kafir	12.9
	Durra-Caudatum	11.9
	Guinea-Bicolor	10.1
	Guinea-Caudatum	6.9
	Guinea-Durra	4.5
	Kafir-Caudatum	3.8
	Kafir-Bicolor	3.8
	Durra-Bicolor	6.9
	Kafir-Durra	1
	Caudatum-Bicolor	0.7
Flowering time	Early flowering < 70 days	3.5
	Mid flowering 70–90 days	78
	Late flowering > 90 days	18.5

color classes present whereas the Kigezi highland region was represented only by red and brown-seeded accessions (Supplementary Table S3).

Grain covering percentage

Glume coverage, or the proportion of glume enclosing the grains showed a high variability in the core collection, which was classified into five categories (Table 4). The majority of accessions (65.7%) had 25% the grain covered by the glume. The least number of accessions mainly from the weedy complexes (hybrids between cultivated and wild sorghum) either had grain fully covered by the glumes (1.05%) or the glumes were longer than the grain (1.05%).

Flowering time

The flowering time (days to 50% flowering) showed a wide distribution ranging from 62 to 132 days. (Supplementary Table S3). The mid-flowering accessions (71–90 days) accounted for 78% of the core collection whereas the early-flowering accessions (< 70 days) were the least frequent (3.5%). The early flowering accessions were mainly from the Southwestern Kigezi highlands and the Northern region, which contributed 60% and 40% of the accessions in this category, respectively. Northern Uganda showed the largest range of flowering time from 62 to 132 days. Accessions from the Northern region greatly dominated the late maturing accessions with about 77%.

Discussion

Core collection composition

It is costly and logistically a huge task to maintain and extensively evaluate a collection of 3333 sorghum accessions in the Uganda National GeneBank. The resources available for evaluating a constantly expanding germplasm collection are limited and steadily decreasing. This calls for the formation of a minimally redundant core collection that captures the maximum genetic diversity in the whole sorghum collection in the national genebank. This study proposes the first *S. bicolor* core collection in the Uganda National GeneBank which reflects the maximum genetic diversity available in the entire collection. GenoCore (Jeong et al. 2017) was used to assemble a core collection of 310 entries which is about 10% of the whole collection. The choice of GenoCore was based on its ability to capture the

maximum number of alleles within the whole collection, which is ideal for germplasm conservation (Schoen and Brown 1993). This proportion of the core collection relative to the whole collection is in line with the recommended size for a good core collection of 5–30% (Brown 1989a, b; van Hintum et al. 2000; Bhattacharjee et al. 2007; Ruiz et al. 2013). This selection is large enough to capture the genetic variability of the available germplasm with a manageable number of accessions.

The core collection well represented the genetic diversity of *S. bicolor* in the Uganda National GenBank in terms of geographical origin, ecology, biological status and ethno-cultural diversity, making it a reliable active collection for implementing ex-situ conservation measures and useable by breeding programs. According to (Brown 1989a, b), a good core collection should have no redundant entries, it should be representative of the whole collection with regards to species, subspecies and geographical regions, and should be small enough to derive reliable conclusions about the whole collection. Representative coverage of diversity is essential because sorghum growing regions possess very diverse environmental conditions in terms of climate, altitude and soil characteristics. Adaptation of Uganda's sorghum landraces to different agro-ecological conditions makes the collection a potential source for favorable alleles for stress acclimation, which are much needed to address the effects of climate change on crop productivity.

Northern and Eastern Uganda contributed about 40% of accessions in the core collection, indicating a high genetic diversity in these regions which was also confirmed by genetic diversity indices. The variation in the core collection composition could be explained by differences in genetic diversity between germplasm from different geographical areas. For instance, accessions from the north are genetically more diverse than accessions from the Kigezi highlands. This is not surprising because the North and West Nile subpopulations are dominated by the guinea race, which is known to be the most genetically diverse race (Menkir et al. 1997; Folkertsma et al. 2005; Bhosale et al. 2011; Billot et al. 2013; Kitavi et al. 2014; Cuevas et al. 2018), whereas the highland region is dominated by durra and caudatum races.

Agro-morphological diversity of Uganda's *Sorghum bicolor* core collection

The core collection captured all the five major sorghum races and their ten intermediate hybrids (Harlan and de Wet 1972; de Wet et al. 1972), which confirms the earlier report by (Reddy et al. 2002) that all the five major sorghum races and their ten intermediates were endemic to Uganda. The assembled sorghum core collection in the Uganda National GenBank is a great opportunity for the national sorghum breeding program to diversify its breeding material by prioritizing specific farmer preferred races in specific agro-ecological zones of Uganda, such as guinea in the north, durra in the semi-arid Karamoja region and caudatum in the Teso and Kigezi regions. For example, the guinea race is currently not utilized by the sorghum breeding programs in Uganda, although it is known to be the most genetically diverse among the cultivated races (Menkir et al. 1997; Folkertsma et al. 2005; Bhosale et al. 2011; Billot et al. 2013; Kitavi et al. 2014; Cuevas et al. 2018). If incorporated in the breeding programs, it can potentially contribute to current and future sorghum breeding efforts targeting West Nile, Acholi and Lango subregions in northern Uganda, where it is the preferred race by farmers. Similarly, durra accessions could be used in breeding new varieties for the drier Karamoja region, as it is the preferred race by farmers in this region. Durra has been reported to thrive in the more arid conditions (Dahlberg 1995, Vadez et al. 2011). This concept of targeted breeding is in tandem with breeding of traditional cereals such as sorghum, which require a decentralized breeding program targeting specific agro-ecological zones with specific farmer preferred landraces for their adaptation, taste or post-harvest processing traits (Ceccarelli and Grando 2007). Trends in sorghum genetic enhancement have shown that targeted varietal release can result in increased adoption of new varieties by farmers (Chintu et al. 1996; Mangombe and Mushonga 1996).

The prominence of guinea accessions in the core collection was not surprising because guinea is the dominant cultivated race in northern Uganda, as in other areas of Southern and Eastern Africa (Folkertsma et al. 2005; Lacy et al. 2006). East Africa is considered to be a secondary center of diversity for guinea (Harlan 1972; Harlan and de Wet 1972; Toure and Scheuring 1982; Barro-Kondombo et al. 2010),

which is considered to be the oldest of the 5 races because of its relatively wide geographical distribution (Harlan and de Wet 1972; deWet et al. 1972). It is highly preferred by farmers in the northern region of Uganda due to its hard corneous grain, pendulous panicles and wide glume opening contributing to resistance to rotting under wet and humid environments (Harlan and de Wet 1972; Haussmann et al. 2012).

The high seed color diversity in the core collection, with seven seed color classes and very high intra-class variations presents an opportunity for breeders to generate specialty sorghum lines that are rich in health-promoting bioactive compounds. Pigmented sorghum is a rich source of antioxidants like anthocyanins and phenolic compounds which have multiple human health benefits (Dicko et al. 2006; Dykes et al. 2014). Bioactive compounds in pigmented sorghum also play a key role in protection against grain mold (Esele et al. 1993) or bird and insect predation (McMillian et al. 1972), although they also impact seed dormancy (Debeaujon et al. 2000). As human nutrition interests are shifting towards maintaining or increasing healthy promoting phytochemicals in grain, it can be expected that the assembled core collection will be an important genetic resource for breeding sorghum varieties with reduced disease-derived phytotoxins and increased health-promoting compounds. Davina et al. (2014) reported that Uganda's sorghum was a good source of germplasm for breeding high polyphenol sorghum.

Variation in sorghum seed color has been attributed to deliberate artificial selection related to grain utilization by the local communities. For example, in Uganda, white grain sorghum is used as food and in commercial beer production, whereas red or brown grain sorghum is used for brewing of traditional alcoholic beverages. This explains the absence of white sorghum in the cold Kigezi highlands, where sorghum is solely used for brewing the traditional alcoholic beverage 'muramba' from darker grain. In Africa, where pests and diseases are common, tannin-containing sorghums are still grown in significant quantities, since they are more tolerant than the non-tannin varieties (Awika and Rooney 2004). This could explain the dominance of red-seeded accessions in the core collection. As suggested by Wu et al. (2012), it is believed that natural selection has retained a certain tannin content in domesticated sorghum, as these compounds conferred sorghum resistance to frequent grain molds and bird damages.

The variation in the proportion of the grain covered by the glume is associated with threshability (Verma et al. 2017) and morphological adaptations facilitating the rapid grain drying process with a minimal risk of grain mold (Gebrie et al. 2019). Therefore, it is not surprising that accessions with grain covering 25% dominated the core set. According to (Upadhyaya et al. 2010), glume cover and color can be utilized to screen for grain mold resistance.

Genetic diversity in the core collection

Northern Uganda showed the greatest diversity for all the scored agro-morphological traits compared to the Eastern and South-western (highland) regions. This could be attributed to the predominance of the guinea race in the region, which has been reported to possess greater genetic diversity among the cultivated races (Menkir et al. 1997; Folkertsma et al. 2005; Bhosale et al. 2011; Billot et al. 2013; Kitavi et al. 2014; Cuevas et al. 2018). The high sorghum diversity in Northern Uganda could be attributed to its location which is adjacent to the southern belt of South Sudan and Ethiopia, a key primary center of sorghum diversity and domestication (Kimber 2000; Mukuru 1993). Northern Uganda is also characterized by a high diversity of landraces, weedy complexes and wild sorghum, including *Sorghum* × *drummondii* (Steud.) Nees ex Millsp. & Chase, *Sorghum purpureosericeum* (Hochst. ex A. Rich.) Schweinf. & Asch., *Sorghum halepense* (L.) Pers., and *Sorghum bicolor* (L.) Moench subsp. *Verticilliflorum* (Steud.) de Wet ex Wiersema & J. Dahlb. (http://www.pgrc.go.ug). Germplasm from this region has considerable potential for improving adaptation to a wide range of environments, compared to the cold Kigezi highland sorghum that is adapted to a specific ecosystem.

The presence of highland sorghum in the core collection represents a unique opportunity for breeders targeting regions with temperate climate. Compared to the other regions, accessions from the cold Kigezi highlands formed a distinct cluster in the PCA (Fig. 1). This could be attributed to the isolation of this region from the lowland regions and its unique climatic conditions (cold stress) may have played a major role in the differentiation of the germplasm from this region. The need for adaptation to cold

climate in this region suggests the presence of potentially unique sorghum genetic resources in the cold Kigezi highlands. Sources for cold stress tolerance have been identified in Uganda's highland sorghum and are being used in sorghum breeding for temperate regions around the world (Johnson and Singh 1975; Singh 1977). The spread and diversification of crops in different locations can lead to new variants, a process influenced by genotype by environment interactions and geographical isolation (Sánchez et al. 2000; Pressoir and Berthaud 2004). However, this process takes time and requires high diversification in ecosystems and genetic isolation. In fact, sorghum has been grown for centuries in the cold Kigezi highland areas of Uganda. These conditions could have differentiated highland sorghum in Southwestern Uganda from the lowland sorghum in the Northern and Eastern regions, due to the significant differences in cultivation environment and infrequent exchange of seeds between the lowland and highland farmers. An intensive analysis of these two groups (lowland and highland sorghum) in the core collection could unveil novel alleles for climatic adaptation.

Sorghum weedy complexes (wild x cultivated sorghum hybrids)

The 18 sorghum weedy complexes in the core collection are most likely the result of wild sorghum x guinea race hybridization events in Northern Uganda. This can be attributed to the open panicles and the long rachis of guinea accessions, which generally leads to a higher frequency of cross pollination (Barnaud et al. 2008). In Northern Uganda, wild and cultivated sorghum commonly occur sympatrically with overlapping flowering phenology, thus potentially allowing gene flow between the two taxa. The observation that wild and cultivated sorghum are inter-fertile and grow in sympatry in sub-Saharan Africa has been documented (Doggett and Majisu 1968; Doggett 1988; Doggett and Prasada Rao 1995; Barnaud et al. 2007; Tesso et al. 2008). The new gene combinations from such events play a key role in the evolution of domesticated species (Slatkin 1987) and continue to increase the genetic diversity in modern crops (Jarvis and Hodgkin 1999). The abundance of sorghum weedy complexes in fields in Northern Uganda is clear evidence that spontaneous hybridization between wild and cultivated sorghum is a common phenomenon in this region. The accession from weedy complexes in the core collection therefore represent an important genetic reservoir for resistance and adaptation traits in sorghum breeding programs (Rooney and Smith 2000; Rosenow and Dahlberg 2000; Bapat and Mote 1982; Karunakar et al. 1994; Franzmann and Hardy 1996; Sharma and Fransmann 2001; Kamala et al. 2002; Komolong et al. 2002, Gurney et al. 2002; Reed et al. 2002; Rao Kameswara et al. 2003; Rich et al. 2004).

Significance of Uganda's *Sorghum bicolor* core collection

Although the sorghum core collections maintained by other countries and the Consultative Group for International Agricultural Research (CGIAR) gene banks represent much of the diversity across the world (Prasada Rao and Ramanatha Rao 1995; Grenier et al. 2000, Grenier et al. 2001; Dahlberg et al. 2004; Deu et al. 2006; Upadhyaya et al. 2007; Shehzhad et al. 2009; Billot et al. 2013), constituent lines of these collections may not be adapted to specific local climatic conditions in Uganda. Therefore, the Uganda national *Sorghum bicolor* core collection is critical for future sorghum breeding in Uganda.

Practically, the core collection will ease the gene bank's activities such as seed regeneration and increases, enabling efficient germplasm exchanges. Similarly, the use of a core collection will simplify the detailed phenotyping and genetic dissection of the gene bank's collection in multi-location trials which is essential for conducting genome-wide association analyses and genomic selection. This could likely enhance the germplasm utilization by enabling the prediction of traits for the non-phenotyped accessions which may carry interesting diversity for traits of interest. Overall, implementation of the core collection will reduce the management costs in the Uganda National GeneBank and avoid unnecessary distribution of genetically related accessions or even duplicates to stakeholders such as breeders.

We expect that the proposed core collection will also stimulate interest, cooperation and coordination and enhance interactions and connections among sorghum geneticists, breeders and other scientists in Uganda and other countries. To ensure an effective conservation and utilization of Uganda's core collection, systematic characterization and proper

documentation is a prerequisite. The core collection database will be uploaded onto the Uganda National GeneBank website and accessed through the Multilateral System (MLS) to allow exchange of germplasm including passport data. Uganda is a signatory to the International Treaty of Plant Genetic Resources for Food and Agriculture (ITPGRFA). Hence, the provisions of the Treaty will be used to exchange the core collection entries subject to existing national legislation. Germplasm conservation is a dynamic process; thus knowledge of the gene pool is never complete and must be continuously improved. In future, the core collection can be updated to include new sorghum accessions shown to have significant new variants that are absent in the present core panel.

Conclusion

The proposed *Sorghum bicolor* core collection of 310 accessions captures the maximum genetic diversity in the whole collection of 3333 accessions maintained by the Uganda National GeneBank. Hence, it qualifies to serve as a reference panel from which useful information will be generated and used as a guide for efficient use of the whole collection. The core collection is currently being evaluated in different agro-ecological zones of Uganda to characterize a number of agronomic traits. Each accession of the core collection has been multiplied and seeds deposited in the Uganda National GeneBank in Entebbe are available upon request according to the ITPGRFA procedures.

Author contribution RM, RJS and LTO conceived the idea, SMW phenotyped the core collection, SC and RM performed the genotype analysis, MF provided statistical advice, RM, LTO, NM, JWM and YB generated the core collection, RM drafted the manuscript, all co-authors contributed to the revision of the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was funded by grant number 393730107 to RJS from the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG).

Data availability The raw DArTseq sequence data generated and analysed during this study has been deposited to the NCBI short-read archive under project number PRJNA779225. Sequence variants reported by DArTseq are available at https://doi.org/10.5281/zenodo.6535431. Phenotype data is available at https://doi.org/10.5281/zenodo.6609823. Seed samples of the core collection are available from the Uganda National Genebank in Entebbe (https://www.pgrc.go.ug/index.php/contactuspgrc) under the Standard Material Transfer Agreement (SMTA) of the United Nations Food and Agriculture Organisation (see https://www.fao.org/plant-treaty/areas-of-work/the-multilateral-system/the-smta/en/).

Declarations

Conflict of interest All authors declare that they have no financial interests related to this work.

References

Awika JM, Rooney LW (2004) Sorghum phytochemicals and their potential impact on human health. Phytochemistry 65(9):1199–1221

Balfourier F, Roussel V, Strelchenko P, Exbrayat-Vinson F, Sourdille P, Boutet G et al (2007) A worldwide bread wheat core collection arrayed in a 384-well plate. Theor Appl Genet 11:1265–1275

Bapat DR, Mote UN (1982) Sources of shoot fly resistance in sorghum. J Maharashtra Agric Univ 7:238–240

Barnaud A, Lacombe TT, Doligez A (2006) Linkage disequilibrium in cultivated grapevine *Vitis Vinifera* L. Theor Appl Genet 112:708–716

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

Barnaud A, Trigueros G, McKey D, Joly HI (2008) High outcrossing rates in fields with mixed sorghum landraces: how are landraces maintained? Heredity 101:445–452

Barro-Kondombo C, Sagnard F, Chantereau J, Deu M, vom Brocke K, Durand P et al (2010) Genetic structure among sorghum landraces as revealed by morphological variation and microsatellite markers in three agroclimatic regions of Burkina Faso. Theor Appl Genet 120:1511–1523

Bhattacharjee R, Khairwal I, Bramel PJ, Reddy K (2007) Establishment of a pearl millet [*Pennisetum glaucum* (L.) R. Br.] core collection based on geographical distribution and quantitative traits. Euphytica 155:35–45

Bhosale SU, Stich B, Rattunde HFW, Weltzien E, Haussmann BIG, Hash CT et al (2011) Population structure in

sorghum accessions from West Africa differing in race and maturity class. Genetica 139(4):453–463
Billot C, Ramu P, Bouchet S, Chantereau J, Deu M, Gardes L et al (2013) Massive sorghum collection genotyped with SSR markers to enhance use of global genetic resources. PLoS ONE 8:e59714. https://doi.org/10.1371/journal.pone.0059714
Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y et al (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635
Brown AHD (1989a) Core collections: a practical approach to genetic resources management. Genome 31:818–824
Brown AHD (1989b) The case for core collections. In: Brown AHD et al (eds) The use of plant genetic resources. Cambridge Univ. Press, Cambridge, England, pp 136–155
Browning BL, Browning SR (2013) Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics 194(2):459–471
Ceccarelli S, Grando S (2007) Decentralized-participatory plant breeding: an example of demand-driven research. Euphytica 155:349–360
Chintu EM, Chigwe CFB, Obilana AB, Chirwa RW, Msiska FS (1996) Sorghum variety release in Malawi: the case of Pirira 1 and Pirira 2. In: Leuschner K, Manthe CS (eds) Drought-tolerant crops of Southern Africa. Proceedings of the SADC/ICRISAT Regional Sorghum and Pearl Millet workshop, 25–29 July 1994, Gaborone, Botswana. International crops research institute for the semi- arid tropics, Patancheru 502 324, Alidhra Pradesh, p 19–25
Cubry P, de Bellis F, Avia K, Bouchet S, Pot D, Dufour M et al (2013) An initial assessment of linkage disequilibrium (LD) in coffee trees: LD patterns in groups of *Coffea canephora* Pierre using microsatellite analysis. BMC Genom 14(1):1–15
Cuevas HE, Prom LK, Cooper EA, Knoll JE, Ni X (2018) Genome-wide association mapping of anthracnose (Colletotrichum sublineolum) resistance in the US sorghum association panel. Plant Genome 11(10):3835
Dahlberg JA, Spinks MS (1995) Current status of the US sorghum germplasm collection. Int Sorghum Millets News Lett 36:4–12
Dahlberg JA, Burke JJ, Rosenow DT (2004) Development of sorghum core collection: refinement and evaluation of a subset from Sudan. Econ Bot 58:556–557
Damon EG (1962) The cultivated sorghums of Ethiopia. Ethiopia College of Agriculture Mech Arts Expt Station Bull 6
Davina HR, Hoffmann L Jr, Rooney WL, Ramu P, Morris GP (2014) Genome-wide association study of grain polyphenol concentrations in global sorghum [sorghum bicolor (L.) moench] germplasm. J Agric Food Chem 62:10916–10927
De Beukelaer H, Smykal P, Davenport GF, Fack V (2012) Core hunter II: fast core subset selection based on multiple genetic diversity measures using mixed replica search. BMC Bioinform 13:312
de Wet JMJ, Harlan JR, Kurmarohita B (1972) Origin and evolution of guinea sorghums. E Afr Agric For J 37:114–119
Debeaujon I, Leon-Kloosterziel KM, Koornneef M (2000) (2000) Influence of the testa on seed dormancy, germination, and longevity in arabidopsis. Plant Physiol 122:403–414
Deu M, Rattunde F, Chantereau J (2006) A global view of genetic diversity in cultivated sorghums using core collection. Genome 49:168–180
Dicko MH, Gruppen H, Traoré AS, Voragen AGJ, van Berkel WJH (2006) Sorghum as human food in Africa: relevance of content of starch and amylase activities. Afr J Biotechnol 5(5):384–395
Doggett H (1988) Sorghum: Longman Scientific and Technical, Burnt Mill, Harlow, Essex, England. Wiley, New York
Doggett H, Majisu BN (1968) Disruptive selection in crop development. Heredity 23:1–26
Doggett H, Prasada KE, R (1995) Sorghum. In: Smartt J, Simmonds NW (eds) Evolution of crop plants, 2nd edn. Longman Group UK limited, London, pp 173–180
Doggett H (1965) The development of cultivated sorghums. In: Hutchinson J (ed) Crop plant evolution. Cambridge Univ Press, London
Dykes L, Hoffmann L Jr, Portillo-Rodriguez D, Rooney WL, Rooney LW (2014) Prediction of total phenols, condensed tannins, and 3-deoxyanthocyanidins in sorghum grain using near-infrared (NOR) spectroscopy. J Cereal Sci 60(1):138–142
El Bakkali A, Haouane H, Moukhli A, Costes E, Van Damme P, Khadari B (2013) Construction of core collections suitable for association mapping to optimize use of mediterranean olive (olea europaea L.) genetic resources. PLoS ONE 8(5):e61265. https://doi.org/10.1371/journal.pone.0061265
Esele JP, Frederiksen RA, Miller FR (1993) The association of genes controlling caryopsis traits with grain mold resistance in sorghum. Phytopathology 83:490
Folkertsma RT, Rattunde HFW, Chandra S, Raju GS, Hash CT (2005) The pattern of genetic diversity of guinea-race *Sorghum bicolor* (L.) Moench landraces as revealed with SSR markers. Theor Appl Genet 111:399–409
Frankel OH (1984) Genetic perspective of germplasm conservation. In: Arber WK, Llimensee K, Peacock WJ, Stalinger P (eds) Genetic manipulation: impact on man and society. Cambridge University Press, Cambridge, pp 161–170
Frankel OH, Brown AHD (1984) Plant genetic resources today: a critical appraisal. In: Holden JHW and Williams JT (eds) Crop genetic resources: conservation & evaluation, George Alien & Unwin Ltd, London, pp 249–257
Franzmann BA, Hardy AT (1996) Testing the host status of Australian indigenous sorghums for the sorghum midge. In: Foale MA, Henzell RG and Kneip JF (eds) Proceedings of the Third Australian Sorghum Conference. Tamworth, NSW, Australia. Melbourne: Australian Institute of Agricultural Science, pp. 365–367
Gebrie G, Genet T (2019) Morphological characterization and evaluation of sorghum [sorghum bicolor (L.) Moench] landraces in Benishangul Gumuz North-western Ethiopia. Greener J Agric Sci 9(1):37–56
Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q et al (2014) TASSEL-GBS: a high-capacity genotyping by sequencing analysis pipeline. PLoS ONE 9(2):e90346
Gouesnard B, Bataillon TM, Decoux G, Rozale C, Schoen DJ, David JL (2001) MSTRAT: an algorithm for building

germplasm core collections by maximizing allelic or phenotypic richness. J Hered 92:93–94
Granato I, Fritsche-Neto R (2017) Snpready: a helper tool to run genomic analysis (r-project.org)
Grenier C, Bramel-Cox PJ, Noirot M, Prasada Rao KE, Hamon P (2000) Assessment of genetic diversity in three subsets constituted from the ICRISAT sorghum collection using random vs. non-random sampling procedures. A. using morpho-agronomical and passport data. Theor Appl Genet 101:190–196
Grenier C, Hamon P, Bramel-Cox PJ (2001) Core collection of sorghum: I. stratification based on eco-geographical data. Crop Sci 41:234–240
Gurney AL, Press MC, Scholes JD (2002) Can wild relatives of sorghum provide new sources of resistance or tolerance against *Striga* species? Weed Res 42:317–324
Harlan JR (1972) A new classification of cultivated sorghum. In: Rao NGP, House LR (eds) Sorghum in the Seventies. Oxford & IBH, New Delhi, pp 512–516
Harlan JR, de Wet JMJ (1972) A simplified classification of cultivated sorghum. Crop Sci 12(2):172–176
Haussmann BIG, Rattunde HF, Weltzien-Rattunde E, Traoré PSC, vom Brocke K, Parzies HK (2012) Breeding strategies for adaptation of pearl millet and sorghum to climate variability and change in West Africa. J Agron Crop Sci 198:327–339
Jarvis D, Hodgkin T (1999) Fanner decision-making and genetic diversity: linking multi-disciplinary research to implementation on farm. In: Bush S (ed) Genes in the field: issues in conserving crop diversity on farm. International Plant Genetic Resources Institute, Rome
Jeong S, Kim JY, Jeong SC, Kang ST, Moon JK, Kim N (2017) GenoCore: a simple and fast algorithm for core subset selection from large genotype datasets. PLoS ONE 12(7):e0181420. https://doi.org/10.1371/journal.pone.0181420
Johnson EC, Singh SP (1975) The development of cool temperature tolerant grain sorghum. In: International Sorghum workshop proceedings. (University of Puerto Rico, Mayaguez Campus) pp 483–495
Kamala V, Singh SD, Bramel PJ, Rao DM (2002) Sources of resistance to downy mildew in wild and weedy sorghums. Crop Sci 42:1357–1360
Kameswara Rao N, Reddy LJ, Bramel PJ (2003) Potential of wild species for genetic enhancement of some semi-arid food crops. Genet Resour Crop Evol 50(7):707–721
Kamvar ZN, Tabima JF, Grünwald NJ (2014) Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. Peer J 4(2):e281
Karunakar RI, Narayana YD, Pande S, Mughogho LK, Singh SD (1994) Evaluation of wild and weedy sorghums for downy mildew resistance. Int Sorghum Millets Newsl 35:104–106
Kim KW, Chung HK, Cho GT, Ma KH, Chandrabalan D, Gwag JG et al (2007) Powercore: a program applying the advanced M strategy with a heuristic search for establishing core sets. Bioinformatics 23(16):515–526
Kimber C (2000) Origins of domesticated sorghum and its early diffusion to India and China. In: Smith CW, Frederickson RA (eds) Sorghum. Wiley, New York, pp 3–98
Kitavi MN, Kiambi DK, Haussman B, Semagn K, Muluvi G, Kairichi M et al (2014) Assessment of the genetic diversity and pattern of relationship of West African sorghum accessions using microsatellite markers. Afr J Biotech 13(14):1503–1514
Komolong B, Chakraborty S, Ryley M, Yates D (2002) Identify and genetic diversity of the sorghum ergot pathogen in Australia. Aust J Agric Res 53:621–628
Lacy SM, Cleveland DA, Soleri D (2006) Farmer choice of sorghum varieties in southern Mali. Hum Ecol 34(3):331–353
Le Cunff L, Fournier-Level1 A, Laucou V, Vezzulli S, Lacombe T, Adam-Blondon AF et al (2008) Construction of nested genetic core collections to optimize the exploitation of natural diversity in Vitis vinifera L. subsp. Sativa. BMC Plant Biol 8: 31
Leroy T, De Bellis F, Legnate H, Musoli P, Kalonji A, Loor Solórzano RG et al (2014) Developing core collections to optimize the management and the exploitation of diversity of the coffee *Coffea* canephora. Genetica 142:185–199. https://doi.org/10.1007/s10709-014-9766-5
Mangombe N, Mushonga JN (1996) Sorghum and pearl millet on-farm research work in Zimbabwe. In: Leuschner K, Manthe CS (eds) Drought-tolerant crops of southern Africa. Proceedings of the SADC/ICRISAT Regional Sorghum and Pearl Millet workshop, 25–29 July 1994, Gaborone, Botswana. International Crops Research Institute for the Semi-Arid Tropics, Patancheru 502 324, Andhra Pradesh, Lidia, pp 81–90
McMillian WW, Wiseman BR, Burns RE, Harris HB, Greene GL (1972) Bird resistance in diverse germplasm of sorghum. Agron J 64:821
Menkir A, Goldsbrough P, Ejeta G (1997) RAPD based assessment of genetic diversity in cultivated races of sorghum. Crop Sci 37:564–569
Mukuru SZ (1993) Sorghum and millet in Eastern Africa. In: Byth DE (ed) Sorghum and millet commodities and research environment. ICRISAT, India, pp 55–62
Noirot M, Hamon S, Anthony F (1996) The principal component scoring: a new method of constituting a core collection using quantitative data. Genet Resour Crop Evol 43:1–6
Odong TL, Jansen J, van Eeuwijk FA, van Hintum TJL (2013) Quality of core collections for effective utilization of genetic resources review, discussion and interpretation. Theor Appl Genet 126:289–305
Perrier X, Flori A, Bonnot F (2003) Data analysis methods. In: Hamon P, Seguin M, Perrier X, Glaszmann JCE (eds) Genetic diversity of cultivated tropical plants. Enfield Science Publishers, Montpellier, pp 43–76
Pressoir G, Berthaud J (2004) Population structure and strong divergent selection shape phenotypic diversification in maize landraces. Heredity 92:95–101
Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559–575
Reddy VG, Rao NK, Reddy BVS, Rao KEP (2002) Geographic distribution of basic and intermediate races in the world collection of sorghum germplasm. Int Sorghum Millets Newsl 43:15–17

Reed JD, Ramundo BA, Claflin LF, Tuinstra MR (2002) Analysis of resistance to ergot in sorghum and potential alternate hosts. Crop Sci 42:1135–1138

Rich PJ, Grenier U, Ejeta G (2004) Striga resistance in the wild relatives of sorghum. Crop Sci 44:2221–2229

Richards CM, Volk GM, Reeves PA, Reilley AA, Henk D, Forsline PL et al (2009) Selection of stratified core sets representing wild apple (Malus sieversii). J Am Soc Hort Sci 134:228–235

Rooney WL, Smith CW (2000) Techniques for developing new cultivars. In: Smith CW, Frederiksen RA (eds) Sorghum: origin, history technology, and production. Wiley, New York

Ruiz M, Giraldo P, Royo C, Carrillo JM (2013) Creation and validation of the spanish durum wheat core collection. Crop Sci 53:2530–2537

Sánchez J, Goodman M, Stuber C (2000) Isozymatic and morphological diversity in the races of maize of Mexico. Economic Bot 54:43–59

Schoen DJ, Brown AHD (1993) Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers. Proc Natl Acad Sci USA 90:10623–10627

Sharma HC, Fransmann BA (2001) Host-plant preference and oviposition responses of the sorghum midge, Stenodiplosis sorghicola (Coquillett) (Dipt., Cecidomyiidae) towards wild relatives of sorghum. J Appl Entomol 125:109–114

Shehzad T, Okuizumi H, Kawase M, Okuno K (2009) Development of SSR based sorghum (Sorghum bicolor (L.) Moench) diversity research set and its evaluation by morphological traits. Genet Resour Crop Evol 56:809–827

Slatkin M (1987) Gene flow and the geographic structure of natural populations. Science 236:787–792

Tesso T, Kapran I, Grenier C, Snow A, Sweeney P, Pedersen J et al (2008) The potential for crop-to-wild gene flow in Sorghum in Ethiopia and Niger: a geographic survey. Crop Sci 48:1425–1431

Thachuk C, Crossa J, Franco J, Dreisigacker S, Warburton M, Davenport GF (2009) *Core hunter*: an algorithm for sampling genetic resources based on multiple genetics measures. Bioinformatics 10:243

Touré AB, Scheuring JF (1982) Presence de genes mainteneurs de l'androsterilité cytoplasmique parmi les variétés locales de sorgho au Mali (In French). L'agronomie Trop 37:362–365

Upadhyaya HD, Ortiz R (2001) A mini core subset for capturing diversity and promoting utilization of chickpea genetic resources. Theor Appl Genet 102:1292–1298

Upadhyaya HD, Shivali SB, Bhattacharjee RR, Gowda CLL, Reddy VG, Singh S (2010) Variation for qualitative and quantitative traits and identification of trait-specific sources in new sorghum germplasm. Crop Pasture Sci 61:609–618

Upadhyaya HD, Reddy VG, Gowda CLL, Singh S (2007) A minicore collection of sorghum [Sorghum bicolor (L) Moench] for enhancing utilization of germplasm in crop improvement. In: The ASA-CSSA-SSSA international annual meetings, New Orleans

Vadez V, Krishnamurthy L, Hash C, Upadhyaya H, Borrell A (2011) Yield, transpiration efficiency, and water-use variations and their interrelationships in the sorghum reference collection. Crop Pasture Sci 62(8):645–655

Van Hintum TJ, Brown A, Spillane C (2000) Core collections of plant genetic resources. Bioversity International, Rome, p 48

Verma R, Ranwah BR, Bharti B, Kumar R, Kunwar R, Diwaker A et al (2017) Characterization of sorghum germplasm for various qualitative traits. J Appl Nat Sci 9(2):1002–1007

Wickham H (2016) ggplot2: elegant graphics for data analysis. Springer, Jun 8

Wu Y, Li X, Xiang W, Zhu C, Lin Z, Wu Y et al (2012) Presence of tannins in sorghum grains is conditioned by different natural alleles of Tannin1. Proc Natl Acad Sci USA 109:10281–10286