

## Genetic diversity and relationships among sweet grain sorghum genotypes across agroclimatic zones revealed by SNP markers

Wendmanegda Hermann TONDÉ<sup>1\*</sup>, Salif BERTHÉ<sup>1</sup>, Patrice BALMA<sup>2</sup>, Jacques OUÉDRAOGO<sup>3</sup>, Tégawendé Odette BONKOUNGOU<sup>1</sup>, Nerbéwendé SAWADOGO<sup>1</sup>

<sup>1</sup>Genetics and Plant Breeding Team/ Biosciences Laboratory, Doctoral School of Science and Technology/ Joseph KI-ZERBO University, 03 BP 7021 Ouagadougou 03, Burkina Faso

<sup>2</sup>Institute of Environment and Agricultural Research (INERA) / Plant Production Department/ Plant Genetics and Biotechnology Laboratory, 04 BP 8645 Ouagadougou, Burkina Faso

<sup>3</sup>University Center of Ziniaré / Joseph KI-ZERBO University, 03 BP 7021, Ouagadougou 03, Burkina Faso

\*Corresponding author: [whermanntonde@gmail.com](mailto:whermanntonde@gmail.com)

ORCID ID: <https://orcid.org/0000-0001-6841-0653>

Submitted:  
16/09/2025

Revised:  
24/11/2025

Accepted:  
29/11/2025

**Abstract:** Sweet grain sorghum is exploited in Burkina Faso through the consumption of fresh grains early at the doughy stage. To identify, valorise and conserve the genetic diversity of this agricultural resource, SNPs markers are essential tools. The aim of the present study is to assess the genetic diversity of sweet grain sorghum expressed in a linear DNA sequence, and to identify SNPs. To this effect, 50 sweet grain sorghum genotypes were sequenced using DArTseq genotyping platform. The results identified 4610 polymorphic loci with a major allele frequency  $\leq 95\%$ . The SNPs identified are made up of 55.23% transitions and 44.77% transversions with an average polymorphism information content of 0.26. The expected heterozygosity value  $H_e$  of 0.18 shows moderate genetic diversity in sweet sorghum in Burkina Faso. The results of the PCoA and AMOVA analyses revealed that genetic diversity is more closely linked to botanical race than to the area of origin of the genotypes. In addition, analysis of the population structure identified two homogeneous subpopulations and one admixture subpopulation. Subpopulation 1 contains 13 genotypes of the *Caudatum-Guinea* race, sub-population 2 contains 32 genotypes of the *Caudatum* race and the admixture sub-population contains 5 genotypes that could belong to the *Guinea-Bicolor* race. Genetic differentiation index was higher between subpopulation 2 and the admixture subpopulation ( $F_{st} = 0.26$ ) and lower ( $F_{st} = 0.13$ ) between subpopulations 1 and 2. The results of this study revealed the first SNP linkage map that can be used to identify QTLs for a marker-assisted breeding program in sweet grain sorghum.

**Keywords:** Burkina Faso; DArTseq; factors; *Sorghum bicolor*; population structure.

**Abbreviations:** DART\_Diversity Array Technologie; DARTseq\_Séquence par Diversity Array Technologie; ILRI\_International Livestock Research Institute.

### Introduction

Sorghum [*Sorghum Bicolor* (L.) Moench] is a diploid species ( $2n = 20$ ) with a genome size of 735 Mbp (Dillon et al., 2005). It is one of the most important cereals grown in the world (Chantereau et al., 2013). Indeed, sorghum is mainly produced for livestock feed in developed countries such as the USA and France (Nicolas, 2007). However, it is a staple food in several tropical regions of Asia and Africa (Guèye et al., 2016). Several production constraints, including drought, striga, insect pests, diseases and low-yielding local cultivars, affect the productivity of sorghum. To address these problems, it is important to have knowledge of the genetic variability of a crop for an efficient selection process (Nemera et al., 2022).

In Burkina Faso, sorghum was the leading crop in terms of area sown ( $\approx 1,900,000$  hectares) and second only to maize in terms of quantity of grain produced ( $\approx 1,800,000$  tonnes) during the 2024-2025 cropping seasons (DGESS/MAAH, 2025). In addition, cultivated sorghum contains significant agromorphological variability (Tondé et al., 2023). Depending on the nature of the part most used, there are several types of cultivated sorghum, including sweet grain sorghum, which is very little exploited and valorised on a national scale (Sawadogo, 2015). It is grown by smallholders, particularly in hut fields (Nebié et al., 2012; Sawadogo et al., 2014). This sorghum has several advantages, including its short cycle and carbohydrate-rich grains at the dough stage (Nebié et al., 2012; Sawadogo et al., 2017). Sweet grain sorghum is therefore harvested before other cereals reach maturity. Panicles harvested at the doughy grain stage are shelled and the fresh grains collected are consumed directly by mastication (Sawadogo, 2015). Sweet grain sorghum requires less irrigation and rainfall and requires fewer inputs compared to sugarcane (Tondé et al., 2023). In addition to its food role, the sale of panicles generates

income for producers and sellers (Sawadogo et al., 2017). The stalks and leaves of this sorghum are also used for livestock feed (Tiendrebéogo et al., 2018).

The various research studies undertaken since 2008 on sweet grain sorghum have revealed genetic diversity using morphological, biochemical and molecular microsatellite markers (Sawadogo, 2015; Sawadogo et al., 2017; Tiendrebéogo et al., 2022). SSR microsatellite markers, which are multi-allelic and highly polymorphic, have been the most widely used DNA markers in molecular diversity studies (Sawadogo et al., 2018; Tiendrebéogo et al., 2022). However, diversity analyses based on genome sequencing offer interesting prospects for exploiting the genetic diversity of this species. Indeed, genetic diversity expressed in a linear DNA sequence makes it possible to identify differences in the genome at the scale of a single nucleotide (Mamo et al., 2023) and to detect the genes responsible for traits of agronomic interest (Lakhanpaul, 2006). This could be possible thanks to the construction of a genetic map based on Single-Nucleotide Polymorphism (SNP). The newly developed sequencing technology, Diversity Arrays Technology (DArTseq) has facilitated the large-scale discovery of SNPs across the whole genome of little-known species (Kilian et al., 2012). DArT markers are polymorphic DNA segments that occur at specific sites in the genome, after complexity reduction, and are detected by hybridisation (Kilian et al., 2012; Cruz et al., 2013). These markers have been successfully applied in genetic diversity studies, linkage mapping and genome-wide association studies for the identification of QTLs in sorghum as well as in several crop species (Jaccoud et al., 2001; Egea et al., 2017; Yahaya et al., 2023). The present study, which uses this technology, aims to (i) identify the various point mutations at the nucleotide level in the sweet grain sorghum genome and (ii) determine the level and structuring of the genetic diversity of this plant genetic resource in Burkina Faso.

## Results

### *SNPs identified and variation in genetic diversity parameters*

A total of 4610 polymorphic loci were enumerated after quality control of genotyping data. Based on the various nucleotide substitutions, the four transition types (A↔G or C↔T) and the eight transversion types (A↔C, A↔T, G↔C, G↔T) were identified. Thus, the nucleotide base changes counted were 55.29% transitions versus 44.71% transversions.

Genetic diversity parameters of the plant material determined from all the polymorphic loci (Table 1) revealed the existence of genetic diversity in the plant material studied. The frequency of the most frequent alleles (MAF) of the loci ranged from 0.50 to 0.95, with an average of 0.89. Observed heterozygosity ( $H_o$ ) averaged 0.07, with minimum and maximum values of 0 and 1, respectively. Expected heterozygosity ( $H_e$ ) recorded a minimum value of 0.08 and a maximum value of 0.5, with an average of 0.18. Polymorphism information content (PIC) values ranged from 0.08 for the least polymorphic loci, to 0.38 for the most polymorphic loci, with an average of 0.26. Individuals' fixation index in the total population ( $F_{IT}$ ) had an average value of 0.75 and varied between -1 and 1.

### *Genetic diversity by factors agroclimatic zones origin and botanical races*

The various genetic diversity parameters calculated for the agroclimatic zones origin factor (Table 2) revealed an average major allele frequency of 0.90 for genotypes from the Sahelian and Southern Sudanian zones and 0.86 for the Northern Sudanian zone. The average observed heterozygosity showed the same value of 0.07 for genotypes from the Sahelian and Northern Sudanian zones and 0.05 for the Southern Sudanian zone. As for expected heterozygosity ( $H_e$ ) and polymorphism information content (PIC), values were highest for the Northern Sudanian zone ( $H_e = 0.21$ ; PIC = 0.18), followed by the Sahelian agroclimatic zone ( $H_e = 0.17$ ; PIC = 0.15) and the Southern Sudanian agroclimatic zone ( $H_e = 0.15$ ; PIC = 0.13). The values obtained for  $F_{IS}$ , which measures the reduction in heterozygosity of individuals within each agroclimatic zones, were all positive. This ranged from 0.58, 0.69 to 0.71 for the Sahelian, Northern Sudanian and Southern Sudanian zones respectively.

The various genetic diversity parameters calculated for the different botanical races (Table 2) revealed genetic diversity was higher for Caudatum race genotypes ( $H_o = 0.45$ ;  $H_e = 0.46$ ). For the Caudatum-Guinea race, mean observed heterozygosity ( $H_o$ ) and mean expected heterozygosity ( $H_e$ ) were 0.21 and 0.35, respectively. The two genotypes of Guinea-Bicolor race ( $H_o = 0.22$ ;  $H_e = 0.28$ ) and the three genotypes not assigned to a race ( $H_o = 0.26$ ;  $H_e = 0.31$ ) registered the lowest genetic diversity. As for polymorphism information content (PIC), mean values were 0.41 for the Caudatum race, 0.31 for the Caudatum-Guinea race, 0.26 for the Guinea-Bicolor race and 0.32 for the unassigned genotypes. All three genetic groups showed a positive fixation index ( $F_{IS} > 0$ ).

The genotypes of the Caudatum race and the Caudatum-Guinea race, which are slightly distant ( $F_{ST} = 0.08$ ), are both very distant from the genotypes of the Guinea-Bicolor race ( $F_{ST} = 0.42$ ;  $F_{ST} = 0.34$ ) and those of the unassigned race ( $F_{ST} = 0.32$ ;  $F_{ST} = 0.26$ ). A moderate distance is recorded between the genotypes of the Guinea-Bicolor and unassigned NA race, with a differentiation index of  $F_{ST} = 0.13$ .

### *Principal coordinate analysis of sweet grain sorghum genotypes*

Principal coordinate analysis (PCoA) based on the pairwise genetic distance matrix gives the distribution of genotypes. For the agroclimatic zone factor (Fig 1A). The total amount of genetic variation explained by the first two principal coordinates is 46.51%. The PCoA shows a mixture of genotypes for all agroclimatic zones. For the botanical race factor (Fig 1B), the total amount of genetic variation explained by the first two principal coordinates is 54.36%, i.e. 38.58% for axis 1 and 15.78% for axis 2. PCoA clearly separated Guinea-Bicolor and NA genotypes. It also showed a high level of overlap between Caudatum and Caudatum-Guinea race genotypes.

### *Genetic structuring of sweet grain sorghum genotypes*

#### *Determining the number of subpopulations in the plant material studied*

Results from the STRUCTURE SELECTOR (Fig 2) show a decreasing curve from  $K = 2$  with a value  $\Delta K = 400$ , to  $K = 4$  with a value  $\Delta K = 100$ . This curve becomes constant from  $K = 5$  with a value  $\Delta K = 0$ . Analysis of this curve reveals that sweet grain sorghum genotypes can be structured between two to four subpopulations. The highest  $\Delta K$  value is obtained at  $K = 2$ , indicating two Subpopulations in the plant material studied.

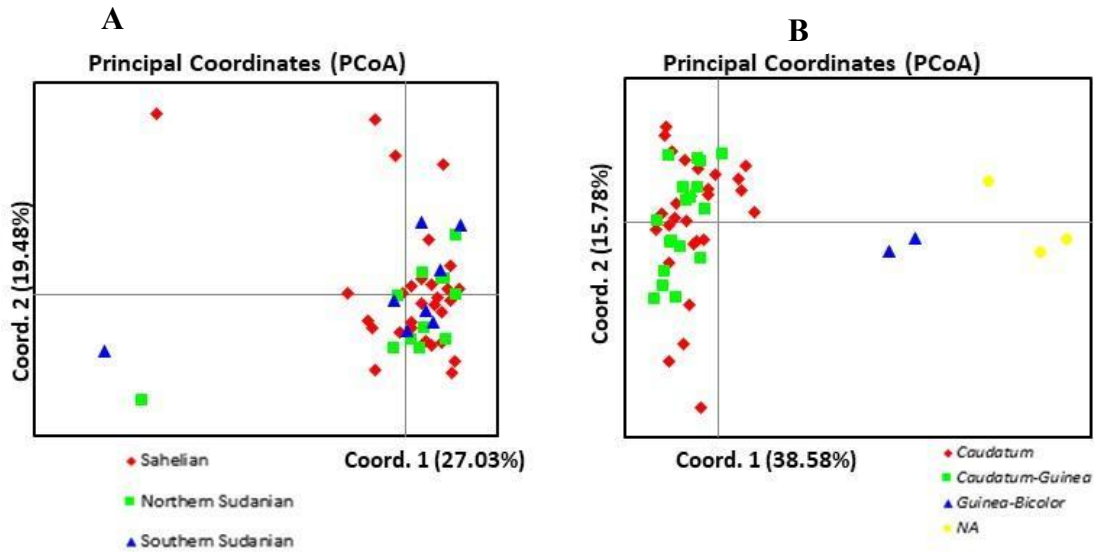
#### *Genotype distribution in subpopulations*

The distribution of genotypes defined by Structure at the highest  $\Delta K$  with a membership coefficient  $Q \geq 80$  (Fig 3) reveals two subpopulations and one admixture subpopulation. Subpopulation 1 comprises 13 genotypes all belonging to the Caudatum-Guinea intermediate race. Subpopulation 2 comprises 32 genotypes, including all 28 genotypes of the main Caudatum race and four genotypes (BZ11, KBA1, YOU5, BK01) of the intermediate Caudatum-Guinea race. The admixture subpopulation contains five genotypes, including

**Table 1.** List of genotypes studied.

Botanical races / Agroclimatic zones	<i>Caudatum</i>	<i>Caudatum-Guinea</i>	<i>Guinea-Bicolor</i>	NA	Total
Sahelian	BK03, GC05, LOU10, LOU2, PGO3, PLA1, SB01, YOH2, YOH7, YOU4	BK01, BZI1, KBA1, LTI3, NBOA1, PBO4, PBO5, PLA3, SKA2, SKA3, SPI2, YOH1, YOH3, YOH8, YOU5	YTA4	SK02, YOH4, YOU1	29
Northern Sudanian	BIP4, BKB2, BKB4, BT02, KNO11, MBO7, MBO8, MDE5, MTC2,	BKB1, KDO12	KDO7	-	12
Southern Sudanian	SBR7, KBZ1, KBZ4, STO2, STO4, STO5, STO6, SBR1, SBR5	-	-	-	09
<b>Total</b>	28	17	02	03	50

Legend: NA: not assigned to a botanical race.

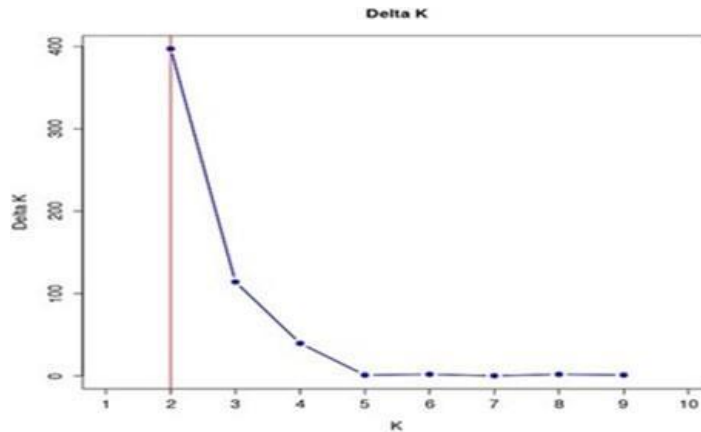


**Fig. 1.** Principal coordinate analysis of sweet grain sorghum genotypes according the agroclimatic zones (A) and botanical races (B).

**Table 2.** Descriptive statistics for genetic diversity indices.

Parameters of genetic diversity	Minimum	Maximum	Mean	P (95%)
MAF	0.50	0.95	0.89	Yes
Ho	0	1	0.06	Yes
He	0.09	0.5	0.18	Yes
PIC	0.08	0.38	0.26	Yes
F <sub>IT</sub>	-1	1	0.75	Yes

Legend: MAF: Major allele frequency; Ho: Observed heterozygosity; He: Expected heterozygosity; PIC: Polymorphism Information Content; F<sub>IT</sub>: Fixation indices of individuals in the total population; P: Probability.



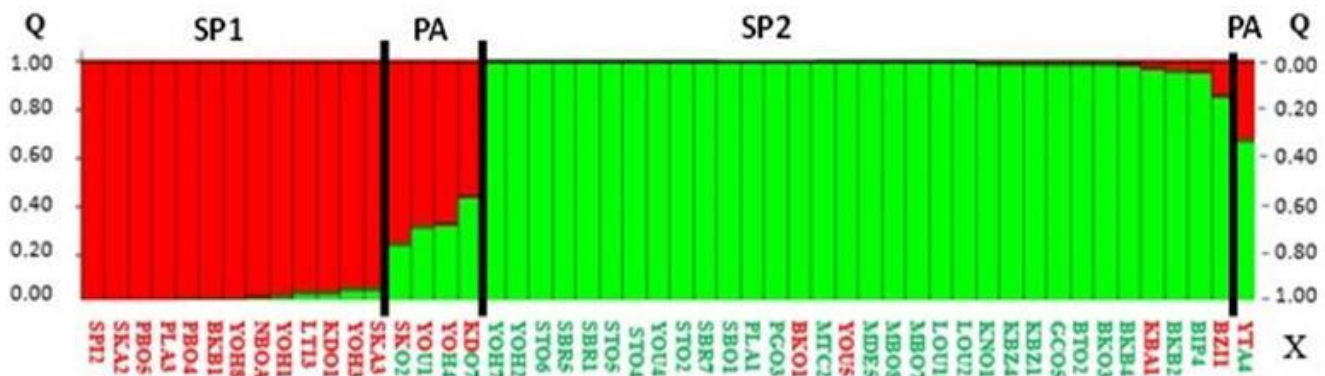
**Fig. 2.** Statistic  $\Delta K$  per number of K subpopulations in the sample.

**Table 3.** Genetic diversity by factors agroclimatic zones of origin and botanical races.

Parameters	MAF	Ho	He	PIC	F <sub>IS</sub>	F <sub>ST</sub>		
						Sahelian	Northern Sudanian	Southern Sudanian
Sahelian	0.9	0.07	0.17	0.15	0.58	0		
Northern Sudanian	0.86	0.07	0.21	0.18	0.69	0.001	0	
Southern Sudanian	0.9	0.05	0.15	0.13	0.71	0.037	0.032	0

Parameters	MAF	Ho	He	PIC	F <sub>IS</sub>	F <sub>ST</sub>			
						Caudatum	Caudatum-Guinea	Guinea-Bicolor	NA
Caudatum	0.66	0.45	0.46	0.41	0.78	0			
Caudatum-Guinea	0.74	0.21	0.35	0.31	0.67	0.08	0		
Guinea-Bicolor	0.48	0.22	0.28	0.26	0.32	0.42	0.34	0	
NA	0.56	0.26	0.31	0.32	0.42	0.32	0.26	0.13	0

**Legend :** MAF : Major allele frequency , He : observed heterozygoty , He : Expected heterozygoty, PIC : Polymorphism Information Content , F<sub>IS</sub> : fixation index of individuals, F<sub>ST</sub> : genetic differentiation index, NA: not assigned to a botanical race



**Fig. 3.** Structure of the 50 sweet grain sorghum genotypes determined at K = 2. **Legend :** X: genotypes; Q: sub-population membership threshold; SP1: subpopulation 1; SP2: subpopulation 2; PA: admixture subpopulation

**Table 4.** Variation of genetic diversity parameters of structure groups and AMOVA.

Genetic diversity parameters								
Subpopulations	MAF	Ho	He	PIC	F <sub>IS</sub>	F <sub>ST</sub>		
						SP1	SP2	PA
SP1	0.74	0.21	0.35	0.31	0.67	0		
SP2	0.82	0.41	0.48	0.51	0.53	0.13	0	
PA	0.58	0.36	0.38	0.32	0.42	0.12	0.26	0

AMOVA						
Source of variation	Df	SQ	MS	VC	PV%	P. value
Among agroclimatic zones	2	8.51	4.26	0.001	0.07	0.522ns
Among botanical races	3	49.35	24.67	1.4	17	0.0012*
Among subpopulations	2	48.25	26.72	1.319	16	0.0022*

**Legend:** SP1: subpopulation 1; SP2: subpopulation 2; PA: admixture subpopulation MAF: major allele frequency; Ho: observed heterozygoty; He: expected heterozygoty; PIC: Polymorphism Information Content; F<sub>IS</sub>: fixation index of individuals in subpopulations, F<sub>ST</sub>: Genetic differentiation index, Df: degrees of freedom; SQ: sum squares; MS: Mean squares VC: variance components; PV: Proportion of Variance; \*p = 0.001; ns: not significant.

two from the Guinea-Bicolor races (YTA4; KDO7) and three genotypes (SKO2; YOH4; YOH1) from race not assigned based on morphological criteria.

**Variation in genetic diversity parameters of structure groups and analyses of molecular variance (AMOVA)**

The various genetic diversity parameters calculated for the structure groups are shown in Table 3. For subpopulation 1, mean observed heterozygoty (Ho) and mean expected heterozygoty (He) were 0.21 and 0.35, respectively. Genetic diversity is higher for subpopulation 2 (Ho = 0.41; He = 0.48) and average for the admixture subpopulation (Ho = 0.36; He= 0.38). As for polymorphism information content (PIC), mean values were 0.31 for subpopulation 1, 0.51 for subpopulation 2 and 0.32 for the admixture subpopulation. All three subpopulations showed positive fixation indexes (F<sub>IS</sub> > 0) with values of 0.67, 0.53 and 0.42, respectively. Genetic differentiation index calculated using the subpopulation factor (Table 4) showed important differentiation between genotypes from subpopulation 2 and the admixture subpopulation (F<sub>ST</sub> = 0.26). It was average between subpopulation 1 and subpopulation 2 genotypes (F<sub>ST</sub> = 0.13) and between subpopulation 1 and the admixture subpopulation recorded (F<sub>ST</sub> = 0.12).

The results of the analyses of molecular variance (AMOVA) showed the contribution of the “botanical race” and “subpopulation” factors in the expression of the molecular variability of plant material (Table 3). The “agroclimatic zone” factor plays a very small role in the expression of variability, with an estimated variance of 0.001 and a contribution to the total variance distribution of 0.072%. Most of the variation is found between botanical races (17%) or between sub-populations (16%) in the expression of total variability. The proportion of variance between botanical races and between subpopulations in the expression of total variability is significant ( $p$ -value < 0.05).

## Discussion

Knowledge of the genetic diversity of cultivated species is a prerequisite for the design of effective breeding and enhancement programs (Santoni et al., 2000). The large number of polymorphic loci obtained in this study indicates variability in the genetic profile of sweet grain sorghum due to point mutations. Indeed, nucleotide substitutions create allelic variations at the same locus, resulting in a difference in the genome of individuals of the same species (Yang et al., 2020). The high percentage of transitions shows that point mutations are more important between purine bases or between pyrimidine bases than between purine and pyrimidine bases. This result could be explained by the different conformations of purine and pyrimidine nuclei. Substitutions between bases on the same nucleus would be easier than between bases on different nuclei. The results of various studies (Vignal et al., 2002) have also found a mutational bias in the favour of transitions. The average polymorphism information content (PIC) value of 0.26 reflects the markers' ability to discriminate between the plant material studied. According to Serrote et al. (2020), PIC is high when its values are between 0.25 and 0.40. However, a molecular evaluation of 120 sweet grain sorghum genotypes found an average PIC value of 0.87 (Sawadogo, 2015). This difference could be explained by the bi-allelic nature of DArT markers, for which the maximum PIC value is 0.5, compared with multi-allelic SSR markers, which have a maximum PIC value of 0.9 (Botstein et al., 1980). Mean value for expected heterozygosity ( $H_e = 0.18$ ) calculated with all loci revealed moderate genetic diversity in the plant material studied. Previous work (Sawadogo, 2015; Tiendrebéogo et al., 2022) found moderate genetic diversity within the collection of sweet grain sorghum genotypes using microsatellite markers. This moderate genetic diversity could be attributed to the selection of desirable traits within the species cultivated by farmers, or to the extent of the collection area. In fact, many authors (Nebié et al., 2012; Sawadogo et al., 2017) have reported that this type of sorghum is endemic to Burkina Faso, where it is grown because of the sweet taste of the grains at the doughy stage. Furthermore, the average observed heterozygosity is lower than the average expected heterozygosity ( $H_o < H_e$ ), i.e. an average fixation index  $F_{IT} > 0$  shows that there are more homozygotes than heterozygotes in the plant material studied. This result could be explained by the mode of reproduction. Sorghum is a monoecious species, preferentially autogamous (Doggett, 1988); due to the minimum  $F_{IT}$  values < 0. The results of the differentiation parameter and PCoA reveal overall mixing of genotypes between agroclimatic zones and proximity of genotypes between botanical races. This result shows that the same genotypes are grown in all three agroclimatic zones. This could be attributed to seed management methods used by growers, such as exchanges or introductions through migratory flows. Thus, the same genotypes can be found in several growing areas in different agroclimatic zones. Previous study has shown that populations take their seeds with them when they migrate and introduce them into host localities (Ouedraogo et al., 2024). This result also indicates that the definition of races is a criterion for classifying sorghum. Indeed, cultivated sorghum has been structured into four main races and eight intermediate races (Harlan and De Wet, 1972) based on panicle type, grain shape and rotation, and the extent of grain coverage by glumes at maturity. The results of the structuring of genetic diversity into two subpopulations are close to the structure into three races, Caudatum and Caudatum-Guinea and Guinea-Bicolor defined based on phenotypic observations by previous author (Sawadogo, 2015). The structuring of molecular and morphological diversity in good agreement confirms that the visible expression of a trait can be under the control of genes. The unassigned race genotypes YOH4 and YOU1, which belong to the admixture subpopulation defined by structure, could be attributed to the multi-lineage character of sorghum or to possible mixtures in the populations; hence the difficulty of phenotypic description. However, the genotypes BZI1, KBA1, YOU5 and BK01 of the Caudatum-Guinea race according to morphological data (Sawadogo, 2015) were distributed in subpopulation 2 with the genotypes of the Caudatum race could belong to this race. The distribution of this genotype in subpopulation 2 could be attributed to designation errors during phenotypic observations or to an environmental influence on the phenotype. Morphological markers, which offer direct phenotypic observation, have the limitation of being influenced by environmental factors (Andersen, 2013). Consequently, molecular tests have been developed to complement conventional selection approaches. Molecular tools remain the most appropriate for assessing genetic diversity, as they are less influenced by environmental variations (Santoni et al., 2000). The higher genetic diversity in subpopulation 2 may be due to the high number of genotypes in this population, or to the botanical type. Subpopulation 2 contains 32 genotypes belonging to the Caudatum race. Furthermore, the analysis of molecular variance in this study, which indicated significant genetic variation between botanical races (17%), could be attributed to the species' preferentially autogamous mode of reproduction. The greater genetic distance between subpopulation 2 and the admixture subpopulation could be mainly due to the botanical race. Indeed, the genotypes of subpopulation 2 are of the Caudatum race and those of the admixture subpopulation are of the Guinea-Bicolor race. The low genetic differentiation index recorded between subpopulations 1 and 2 could be explained by their racial composition. In fact, these two subpopulations are made up, respectively, of genotypes from the Caudatum-Guinea and Caudatum race, and; therefore, have a closer genetic make-up. This result could confirm the possibility of hybridisation between the Guinea and Caudatum genotypes that gave rise to the Caudatum-Guinea intermediate race. Inter-racial crossing could; therefore, be the best way to recombine favourable traits (Trouche et al., 1999) in sweet grain sorghum breeding.

## Materials and Methods

### *Plant material*

The plant material studied consists of 50 sweet grain sorghum genotypes. These genotypes were obtained from the gene bank of the Plant Genetics and Improvement team of the Biosciences Laboratory of the Joseph KI-ZERBO University (Burkina Faso). The genotypes used in this study (Table 1) collected from three agroclimatic zones, namely Sahelian, northern Sudanian and southern Sudanian. They belong to the main Caudatum race and two intermediate races, Caudatum-Guinea and Guinea-Bicolor. However, three genotypes (SK02, YOH4 and YOU1), could not be assigned to a race on the basis of morphological traits (Sawadogo, 2015).

### *Molecular markers used*

Single-nucleotide polymorphism (SNP) markers were used for this study. These markers were generated following sequencing of the sweet grain sorghum genome carried out at the SEQART AFRICA laboratory of the International Livestock Research Institute (ILRI) in

Nairobi, Kenya. The SEQART AFRICA laboratory uses DArTseq™ genotyping-by-sequencing (GBS) technology, which enables rapid, high-quality and affordable genome profiling, even from the most complex polyploid genomes.

### Sequencing methodology

Leaf discs were first collected from two-week-old plants of the different genotypes, then dried in an oven at 70 C for 72 h. Dried leaf discs were then stored in convenient sample collection kits consisting of a storage rack and 96-well tubes. The kits were shipped to the SEQART AFRICA laboratory for genotyping. In the laboratory, DNA extraction was first carried out from these leaf discs using the Nucleomag plant DNA extraction kit. The genomic DNA extracted was between 50 and 100 ng/μL. DNA quality and quantity were checked on 0.8% agarose. Libraries were then constructed using the DArTSeq complexity reduction method (Kilian et al., 2012) by digesting genomic DNA with a combination of PstI and HpaII enzymes and ligating barcoded and common adapters, followed by PCR amplification of the adapter-ligated fragments. Libraries were sequenced using single-read sequencing cycles for 77 bases. Next-generation sequencing was performed using HiSeq2500. Finally, DArTseq marker evaluation was performed using DArTsoft14, an in-house algorithm-based marker evaluation pipeline. Two types of DArTseq markers were scored, SilicoDArT markers and SNP markers, both of which were scored binary for the presence/absence (1 and 0, respectively) of the restriction fragment with the marker sequence in the genomic representation of the sample. SilicoDArT markers and SNP markers were aligned on Sorghum\_v21, to identify chromosomal positions

### Data analysis

#### Data quality control and characterisation of mutation types

The raw sequencing data were filtered based on the amount of missing data and the frequency of occurrence of the major allele (MAF) for each locus (Reed et al., 2015). First, all markers with unidentified chromosomes were discarded. Next, loci with more than 20% missing data were removed. Finally, all loci whose most frequent allele (MAF) has a frequency > 0.95 were also discarded (Chattopadhyay et al., 2014; Linck and Battey, 2019). The number and frequency of mutation types were then determined from the filtered data. Sequencing data filtration and mutation type characterisation were performed using Excel 2019 spreadsheet and Tassel 5.0 software (Bradbury et al., 2007).

#### Estimates of genetic diversity

Genetic diversity was estimated at both intrapopulation and interpopulation levels (Botstein et al., 1980) using Power Marker 3.25 software (Liu and Muse, 2005).

- Observed heterozygosity ( $H_o$ ), which represents the proportion of individuals heterozygous at a given  $k$  locus.  $H_o = 1 - \sum_i^k n_{AiAi}/N$   
 $n_{AiAi}$  is the number of individuals with the  $AiAi$  genotype (Homozygotes),  $N$  is the total number of individuals

- Expected heterozygosity ( $H_e$ ) or Nei's gene diversity, which is the proportion of heterozygous loci for a given population under Hardy-Weinberg equilibrium conditions.

$$H_e = 1 - \frac{1}{m} \sum_{i=1}^m \sum_{j=1}^n f_i^2$$

$f_i$  is the frequency of allele  $i$ ,  $m$  is the total number of loci and  $n$  is the total number of alleles.

- Polymorphism Information Content (PIC), which is the marker's ability to establish polymorphism in the plant material studied, based on the number and frequency of distribution of the alleles identified. Hence,

$$PIC = 1 - \sum_{i=1}^n p_i^2 - \sum_{k=1}^n \sum_{j=l+1}^n 2p_i^2 p_j^2$$

$n$  = number of alleles;  $p_i$  = frequency of the  $i$ th allele;  $p_j$  = frequency of the  $j$ th allele

- Individual fixation index in the total population ( $F_{IT}$ ) or in subpopulations ( $F_{IS}$ ). They measure the deviation from panmixy or the overall deficit in heterozygotes respectively at the overall population level and for subpopulations, i.e.

$$F_{IT} = 1 - \frac{H_o}{H_T}; F_{IS} = 1 - \frac{H_o}{H_S}$$

$H_T$ : the expected proportion of heterozygotes in the total population;  $H_S$ : the expected proportion of heterozygotes in the subpopulations.

- Wright's fixation index ( $F_{st}$ ), measures the heterozygote deficit due to differentiation between subpopulations.

$$F_{st} = 1 - \frac{H_s}{H_t}$$

$H_s$  = intra-population genetic diversity;  $H_t$  = total genetic diversity, where  $H_t = H_s + D_{st}$  and  $D_{st}$  = inter-population genetic diversity.

- Principal Coordinates Analysis (PCoA), a method that uses distances was permitted to analyse and visualise using two-dimension PCA plot, population relationships for these factors. Analysis of molecular variance (AMOVA) was finally used to estimate the proportion of genetic variation for all levels of population structure. PCoA and AMOVA were determined using GenAlex 6.5 software (Peakall and Smouse, 2012).

#### Population structure analysis

Genetic structure of the studied genotypes was analysed according to Bayesian-based approach using the software STRUCTURE version 2.3.4 (Pritchard et al., 2000). This involved generating the number of subpopulations  $K$ , that best describes the organisation of genetic diversity within the plant material studied. The program was run with a burn-in period of 100,000 iterations, further run length of 200,000 Markov Chain Monte Carlo steps, testing population subdivision from  $K = 1$  to 10 under the admixture and correlated allele frequencies model, without prior information on sampling locations (Thapa et al., 2021). The 100 independent simulations were performed for each  $K$  to identify the optimal  $K$  value based on the methods of the maximum likelihood  $L(K)$  (Pritchard et al., 2000) and the ad hoc quantity ( $\Delta K$ ) approaches (Evanno et al., 2005) implemented in the software STRUCTURE Selector (Li and Liu, 2018). This method uses an ad hoc statistic  $\Delta K$ , which considers the rate of change in the log probability of data between successive  $K$  values. Results from STRUCTURE are presented in histogram form, where the different bars correspond to the size of the genomic fraction ( $Q$ ) of the genotypes. Each of the 50 genotypes was attributed to a given subpopulation when the proportion of its genome  $Q \geq 0.80$  or a subpopulation admixture ( $Q < 0.80$ ).

## Conclusion

The molecular markers have become essential tools for identifying genetic diversity of agricultural species. The results of sequencing using the new DArTseq genotyping technology have identified 4610 polymorphic loci induced by point mutations. The diversity estimated for all loci revealed moderate genetic diversity ( $H_e = 0.18$ ), an excess of homozygotes ( $F_{IT} > 0$ ). In addition, analysis of the organization of genetic diversity enabled us to structure the plant material into two subpopulations and one admixture subpopulation, depending on the three botanical races defined based on morphological criteria. The subpopulation 2 subpopulation, which contains 32 genotypes belonging to the Caudatum race highest genetic diversity ( $H_e = 0.48$ ). In addition, the genetic distance was lowest ( $F_{st} = 0.13$ ) between subpopulation 2 and subpopulation 1, made up of genotypes belonging to the Caudatum-Guinea race. The highest differentiation index ( $F_{st} = 0.26$ ) was obtained between subpopulation 2 and the admixture subpopulation, which contain genotypes of the Guinea-Bicolor race. The results of this study offer possibilities for studying the genetic determinism of quantitative traits of agronomic interest.

## Acknowledgements

We express our gratitude to the managers and members of the Biosciences Laboratory of the Joseph KI-ZERBO University for the financial support granted in the sequencing of sweet grain sorghum.

## References

- Andersen SB (2013) Plant Breeding from Laboratories to Fields. BoD – Books on Demand, 300p.
- Botstein D, White RL, Skolnick M, Davis RW (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, 32(3), 314-331.
- Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics*, 23(19): 2633-2635.
- Chantereau J, Cruz JF, Ratnadass A, Trouche G, Fliedel G (2013) *Le sorgho*. éditions Quae.
- Chattopadhyay B, Garg KM, Ramakrishnan U (2014). Effect of diversity and missing data on genetic assignment with RAD-Seq markers. *BMC Research Notes* 7(1): 841.
- Cruz VMV, Kilian A, Dierig DA (2013) Development of DArT Marker Platforms and Genetic Diversity Assessment of the U.S. Collection of the New Oilseed Crop *Lesquerella* and Related Species. *PLOS ONE* 8(5): 1-13.
- Direction Générale des Etudes Statistiques Sectoriels/Ministère de L'agriculture et des Aménagements Hydrauliques (2025) Résultats Définitifs de la Campagne Agricole et de la Situation Alimentaire et Nutritionnelle 2024/2025, Burkina Faso. 39p.
- Dillon SL, Lawrence PK, Henry RJ (2005) The new use of *Sorghum bicolor*-derived SSR markers to evaluate genetic diversity in 17 Australian Sorghum species. *Plant Genetic Resources* 3(1):19-28.
- Doggett H (1988) Sorghum. Harlow, Essex, England: Longman Scientific and Technical, Wiley (New York), 512 p.
- Egea LA, Mérida-García R, Kilian A, Hernandez P, Dorado G (2017) Assessment of Genetic Diversity and Structure of Large Garlic (*Allium sativum*) Germplasm Bank, by Diversity Arrays Technology "Genotyping-by-Sequencing" Platform (DArTseq). *Frontiers in Genetics* 8:1-9.
- Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE : A simulation study. *Molecular Ecology* 14(8): 2611-2620.
- Guèye T, Sine B, Cisse N, Diatta C, Ndiaye S (2016) Characterization of Phenotypic Diversity of Sorghum Collection for Developing Breeding Material. *International Journal of Sciences* 5(02): 38-48.
- Harlan JR, De Wet JMJ (1972) A Simplified Classification of Cultivated Sorghum. *Crop Science* 12(2): 172-176.
- Jaccoud D, Peng K, Feinstein D, Kilian A (2001) Diversity Arrays : A solid state technology for sequence information independent genotyping. *Nucleic Acids Research* 29(4): 1-7.
- Kilian A, Wenzl P, Huttner E, Carling J, Xia L, Blois H, Caig V, Heller-Uszynska K, Jaccoud D, Hopper C, Aschenbrenner-Kilian M, Evers M, Peng K, Cayla C, Hok P, Uszynski G (2012) Diversity arrays technology : A generic genome profiling technology on open platforms. *Methods in Molecular Biology* 888: 67-89.
- Lakhanpaul S (2006) Single nucleotide polymorphism (SNP)–Methods and applications in plant genetics : A review 5: 435-459.
- Li YL, Liu JX (2018) StructureSelector : A web-based software to select and visualize the optimal number of clusters using multiple methods. *Molecular Ecology Resources* 18(1): 176-177.
- Linck E, Battey CJ (2019) Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Molecular Ecology Resources* 19(3): 639-647.
- Liu K, Muse SV (2005) PowerMarker : An integrated analysis environment for genetic marker analysis. *Bioinformatics* 21(9): 2128-2129.
- Mamo W, Enyew M, Mekonnen T, Tesfaye K, Feyissa T (2023) Genetic diversity and population structure of sorghum [*Sorghum bicolor* (L.) Moench] genotypes in Ethiopia as revealed by microsatellite markers. *Heliyon* 9(1): 1-12.
- Nebé B, Gapili N, Traoré RE, Nanema KR, Bationo-Kando P, Sawadogo M, Zongo JD (2012) Diversité phénotypique de sorgho à grains sucrés du centre-nord du Burkina Faso. *Sciences et techniques, sciences naturelles et agronomie* 32(1): 73-84.
- Nemera B, Kebede M, Enyew M, Feyissa T (2022) Genetic diversity and population structure of sorghum [*Sorghum bicolor* (L.) Moench] in Ethiopia as revealed by microsatellite markers. *Acta Agriculturae Scandinavica, Section B-Soil & Plant Science* 72(1): 873-884.
- Nicolas D. (2007) *Utilisation du sorgho en alimentation animale*. Thèse présentée en vue de l'obtention du grade de docteur vétérinaire, Université Claude-Bernard - Lyon I, 109p.
- Ouedraogo J, Kiebre Z, Nanema KR, Bationo/Kando P (2024) Genetic Diversity of Amaranth in Burkina Faso. *Journal of Plant Biotechnology* 51(2024): 1-38.
- Peakall R, Smouse PE (2012) GenAEx 6.5 : Genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics* 28(19): 2537-2539.
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of Population Structure Using Multilocus Genotype Data. *Genetics* 155(2): 945-959.

- Santoni S, Faivre-Rampant P, Prado E, Prat D (2000) Marqueurs moléculaires pour l'analyse des ressources génétiques et l'amélioration des plantes. Cahiers Agricultures 9(4): 311-327.
- Sawadogo N (2015) *Diversité génétique des sorghos à grains sucrés [Sorghum bicolor (L.) Moench] du Burkina Faso*. Thèse Unique, Université de Ouagadougou, Burkina Faso 182 p.
- Sawadogo N, Batiéno TBJ, Kiébré Z, Ouédraogo MH, Zida WMS F, Nanema KR, Nébié, B, Bationo-Kando P, Traoré RE, Sawadogo M, Zongo JD (2018) Assessment of genetic diversity of Burkina Faso sweet grain sorghum using microsatellite markers. African Journal of Biotechnology 17(12): 389-395.
- Sawadogo N, Nanema KR, Bationo/Kando P, Traore RE, Nebie B, Tiama D, Sawadogo M, Zongo JD (2014) Évaluation de la diversité génétique des sorghos à grains sucrés (*Sorghum bicolor* (L.) Moench) du Nord du Burkina Faso. Journal of Applied Biosciences 84(1): 7654-7664.
- Sawadogo N, Ouédraogo MH, Traoré RE, Nanema KR, Kiébré Z, Bationo-Kando P, Nebié B, Sawadogo M, Zongo JD (2017) Effect of agromorphological diversity and botanical race on biochemical composition in sweet grains Sorghum [*Sorghum bicolor* (L.) Moench] of Burkina Faso. Journal of BioScience and Biotechnology 6(1): 263-269.
- Serrote CML, Reiniger LRS, Silva KB, Rabaiolli SMDS, Stefanel CM (2020) Determining the Polymorphism Information Content of a molecular marker. Gene 726(1): 1-14.
- Thapa R, Edwards M, Blair MW (2021) Relationship of Cultivated Grain Amaranth Species and Wild Relative Accessions. Genes, 12(12):1-14.
- Tiendrebéogo J, Sawadogo N, Kiébré M, Kaboré B, Bationo/Kando P, Kiendrebéogo T, Ouédraogo MH, Sawadogo M (2018) Evaluation comparative de la production de grains et du fourrage de sorgho à grains sucrés du Burkina Faso. SPECIAL SIST 2017 SNA2\_agrono27i 11(36): 261-271.
- Tiendrebéogo J, Sawadogo N, Kiébré M, Kiébré Z, Tuina S, Sawadogo TA, Nanema KR, Traoré RE, Ouédraogo MH, Sawadogo M (2022) Genetic Relationship between Sweet Grain Sorghum and the Other Sorghum Types Cultivated in Burkina Faso Assessed with Nuclear Microsatellite Markers SSRs. American Journal of Plant Sciences 13(6): 872-883.
- Tondé WH, Sawadogo N, Tiendrebéogo J, Sawadogo P, Ouédraogo MH, Bougma LA, Sawadogo M (2023) Genetic Diversity, Importance and Prospects for Varietal Improvement of Sweet Grain Sorghum in Burkina Faso. International Journal of Zoology and Applied Biosciences 8(1): 44-52.
- Trouche G, Fliedel G, Chantereau J, Barro C (1999) Productivité et qualité des grains de sorgho pour le tô en Afrique de l'Ouest : Les nouvelles voies d'amélioration. Agriculture et développement 23: 94-107.
- Vignal A, Milan D, SanCristobal M, Eggen A (2002) A review on SNP and other types of molecular markers and their use in animal genetics. Genetics Selection Evolution, 34: 275-305.
- Yahaya MA, Shimelis H, Nebie B, Ojiewo CO, Rathore A, Das R (2023) Genetic Diversity and Population Structure of African Sorghum (*Sorghum bicolor* L. Moench) Accessions Assessed through Single Nucleotide Polymorphisms Markers. Genes 14(7): 1-16.
- Yang S, Gill RA, Zaman QU, Ulhassan Z, Zhou W (2020) Insights on SNP types, detection methods and their utilization in Brassica species : Recent progress and future perspectives. Journal of Biotechnology 324: 11-20.