ISSN 1479-2621

Plant Genetic Resources: Characterization and Utilization; 1–9
doi:10.1017/S1479262121000034

Assessment of diversity in tropical soybean (*Glycine max* (L.) Merr.) varieties and elite breeding lines using single nucleotide polymorphism markers

Abush Tesfaye Abebe[1], Adesike Oladoyin Kolawole[2], Nnanna Unachukwu[1], Godfree Chigeza[3], Hailu Tefera[4] and Melaku Gedil[1]*

[1]*International Institute of Tropical Agriculture, Ibadan, Nigeria,* [2]*Ladoke Akintola University of Technology, Ogbomoso, Nigeria,* [3]*International Institute of Tropical Agriculture, Lusaka, Zambia and* [4]*Private Consultant, 2384 Rolling Fork Circle, #403, Herndon, VA 20171, USA*

Received 26 June 2020, revised 26 January 2021; Accepted 27 January 2021

Abstract

Soybean (*Glycine max* (L.) Merr.) is an important legume crop with high commercial value widely cultivated globally. Thus, the genetic characterization of the existing soybean germplasm will provide useful information for enhanced conservation, improvement and future utilization. This study aimed to assess the extent of genetic diversity of soybean elite breeding lines and varieties developed by the soybean breeding programme of the International Institute of Tropical Agriculture (IITA), Ibadan, Nigeria. The genetic diversity of 65 soybean genotypes was studied using single-nucleotide polymorphism (SNP) markers. The result revealed that 2446 alleles were detected, and the indicators for allelic richness and diversity had good differentiating power in assessing the diversity of the genotypes. The three complementary approaches used in the study grouped the germplasm into three major clusters based on genetic relatedness. The analysis of molecular variance revealed that 71% ($P < 0.001$) variation was due to among individual genotypes, while 11% ($P < 0.001$) was ascribed to differences among the three clusters, and the fixation index (F_{ST}) was 0.11 for the SNP loci, signifying moderate genetic differentiation among the genotypes. The identified private alleles indicate that the soybean germplasm contains diverse variability that is yet to be exploited. The SNP markers revealed high diversity in the studied germplasm and found to be efficient for assessing genetic diversity in the crop. These results provide valuable information that might be utilized for assessing the genetic variability of soybean and other legume crops germplasm by breeding programmes.

Keywords: allelic diversity, breeding programmes, genetic relatedness, SNP, soybean germplasm

Introduction

Soybean (*Glycine max* (L.) Merr.) is one of the major legumes and oil crops of the world, in terms of total production and trade (Chen and Nelson, 2005; Hymowitz and Shurtleff, 2005; FAO, 2020). However, the genetic improvement of such an essential crop has been challenged by its extremely narrow genetic base (Gizlice *et al.*, 1993, 1994, 1996; Salado-Navarro *et al.*, 1993; Sneller, 1994; Singh and Hymowitz, 1999; Cornelious and Sneller, 2002). To sustain its genetic diversity, over 170,000 soybean germplasms have been conserved across 17 countries globally; while the Chinese National Crop Genebank maintains 31,575 accessions (Qiu *et al.*, 2011) and the soybean genetic resource centre of the United States Department of Agriculture (USDA) maintains over 18,000 germplasm collections (Soybase, 2020). These germplasm collections have made significant contributions to production and breeding programmes, since they possess several unique

*Corresponding author. E-mail: m.gedil@cgiar.org

genes that can be utilized for the genetic improvement of the crop (Qiu *et al.*, 2011; Soybase, 2020). Populations derived from the genetic recombination of biparental crosses of diverse parents might be vital sources of higher genetic variability (Helms *et al.*, 1997; Kisha *et al.*, 1997). Some reports show the contributions of improved varieties (e.g. crosses of elite X elite varieties) and elite breeding lines developed by breeding programmes are on the rise, compared to germplasm collections and landraces. For instance, in China, elite breeding lines from crosses and cultivars contributed to 36 and 86% of the soybean varieties released in the 1950s, and between 1993 and 2004, respectively, relative to landraces, traditional varieties and wild relatives (Qiu *et al.*, 2011).

Progress in genetic improvement of any crop depends on the presence of genetic diversity within the populations under selection. Knowledge of the genetic diversity crops is very important for designing strategies to establish core collections and enhance utilization of the germplasm by breeding programmes. Soybean production is gaining increasing importance in several sub-Saharan African (SSA) countries, such as Nigeria, Ghana, Uganda, Ethiopia, Zambia and Malawi (FAO, 2020), the crop being cultivated in the wider agro-ecological conditions of these countries. The soybean improvement programme at the International Institute of Tropical Agriculture (IITA), Ibadan, Nigeria, has a soybean collection of more than 1800 accessions, including high-yielding cultivars without information on the extent of their genetic diversity. These accessions have been phenotypically screened for priority traits that include disease resistance, pod shattering tolerance, early and medium maturity, efficient natural nodulation, lodging tolerance and yield improvement (Tefera *et al.*, 2009; Chigeza *et al.*, 2019). These authors also reported the release of more than 100 IITA bred soybean varieties by the National Agricultural Research Systems across several countries in SSA, following the screening efforts.

Numerous studies have been carried out to determine the degree of genetic variation of varieties and breeding lines of soybean and their relatedness (Keim *et al.*, 1992; Gizlice *et al.*, 1994; Sneller, 1994; Sneller *et al.*, 1997; Kisha *et al.*, 1998; Nelson *et al.*, 1998). Due to its commercial value, soybean has been the subject of advanced genomic studies by the private and public sectors. A number of molecular genomic resources, including single-nucleotide polymorphisms (SNPs), such as: Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html), SoyKB (http://soykb.org/) and Soybase (https://soybase.org/) are available in the public databases. Despite the abundance of genomic resources, no molecular characterization has been performed on the soybean breeding lines at IITA. Therefore, the objective of this study was to assess the extent of genetic diversity of breeding lines and varieties developed by the IITA soybean breeding programme.

Materials and methods

Plant material and DNA isolation

A total of 65 soybean genotypes (17 released varieties and 48 elite breeding lines) from IITA's soybean breeding programme were used in this study (Table 1). Although most of these varieties were released in Nigeria, some lines have been released in Ghana, Benin, Togo, Democratic Republic of the Congo, Uganda, and Ethiopia from 1989 to 2011. All the genotypes with TGx 1987 series are rust-resistant, as they are developed from a cross of rust-resistant donor parent UG-5, in addition to other desirable traits. TGx 1835-10E is a variety released in Nigeria for early maturity and rust resistance. Soy104 and UG-5 were unimproved, but sources of rust resistance genes. TGx 1440-1E and TGx 1448-2E are suitable as a trap varieties in depleting the seed bank of *Striga hermonthica*.

DNA extraction and genotyping

For SNP genotyping, the 65 soybean genotypes were planted and grown into seedlings for 3 weeks, after which fresh bulk young leaves were harvested from all the seedlings (10) per genotype and ground into a fine powder using liquid nitrogen. The genomic DNA of each plant sample was extracted using a miniprep Dellaporta extraction protocol (Dellaporta *et al.*, 1983). The quality of the extracted DNA samples was checked using a 1% agarose gel prepared in 150 ml of 1× TBE agarose gel and quantified using Nanodrop Technologies ND-1000 model range of a nanodrop spectrophotometer.

GoldenGate assay-based SNP genotyping

The genomic DNA extracted from each of the 65 soybean genotypes was sent to the Soybean Genomics and Improvement Laboratory, Beltsville Agricultural Research Center (SGIL-BARC), Beltsville, Maryland. DNA quality was also checked at SGIL-BARC and immediately used for the GoldenGate assay. The GoldenGate assay was performed as per the procedure described by Hyten *et al.* (2008).

Statistical analyses

Genetic diversity indices

SNP markers with a minor allele frequency (MAF) less than 0.05 were filtered out, resulting in 1223 informative loci used in the analyses. The genetic properties of SNPs, such as MAF, polymorphic information content (PIC) and percentage of polymorphic loci (% *P*) were calculated to quantify the genetic diversity within and among the 65 soybean genotypes. In addition, genetic diversity indices, such

Table 1. Rust reaction and pedigree of the soybean genotypes used in the genetic diversity study

Sl. No.	Genotype	Rust reaction[a]	Pedigree	Sl. No.	Genotype	Rust reaction[a]	Pedigree
1	SOY104	Res	USA line	34	TGx 1895-33F	Susc	TGx 1814-3E X TGx1740-6F
2	TGx 1835-10E	Res	TGx 1213-1D X TGx 1445-3E	35	TGx 1903-1F	Susc	TGx 1740-2F X TGx 1830-32E
3	TGx 1987-11F	Res	TGx 1805-31F X UG5	36	TGx 1903-3F	Susc	TGx 1740-2F X TGx 1830-32E
4	TGx 1987-17F	Res	TGx 1805-31F X UG5	37	TGx 1903-7F	Susc	TGx 1740-2F X TGx 1830-32E
5	TGx 1987-18F	Res	TGx 1805-31F X UG5	38	TGx 1903-8F	Susc	TGx 1740-2F X TGx 1830-32E
6	TGx 1987-20F	Res	TGx 1805-31F X UG5	39	TGx 1904-2F	Susc	TGx 1802-2F X TGx 1834-5E
7	TGx 1987-23F	Res	TGx 1805-31F X UG5	40	TGx 1904-6F	Susc	TGx 1802-2F X TGx 1834-5E
8	TGx 1987-25F	Res	TGx 1805-31F X UG5	41	TGx 1908-8F	Susc	TGx 1829-1E X TGx 1834-5E
9	TGx 1987-28F	Res	TGx 1805-31F X UG5	42	TGx 1910-14F	Susc	TGx 1834-5E X TGx 1828-4E
10	TGx 1987-31F	Res	TGx 1805-31F X UG5	43	TGx 1919-1F	Susc	TGx 1485-1D X TGx1830-20E
11	TGx 1987-32F	Res	TGx 1805-31F X UG5	44	TGx 1924-2F	Susc	TGx 1828-3F X TGx 1802-3F
12	TGx 1987-34F	Res	TGx 1805-31F X UG5	45	TGx 1925-1F	Susc	TGx 1848-2E X TGx 1802-3F
13	TGx 1987-62F	Res	TGx 1805-31F X UG5	46	TGx 1935-3F	Susc	TGx 1848-9E X TGx 1848-10E
14	TGx 1987-64F	Res	TGx 1805-31F X UG5	47	TGx 1937-1F	Susc	TGx 1805-31F X TGx 1834-5E
15	TGx 1987-65F	Res	TGx 1805-31F X UG5	48	TGx 1945-1F	Susc	TGx 1848-9E X TGx 1734-5E
16	TGx 1987-6F	Res	TGx 1805-31F X UG5	49	TGx 1949-7F	Susc	TGx 1843-26E X TGx1848-9E
17	TGx 1987-8F	Res	TGx 1805-31F X UG5	50	TGx 1950-7F	Susc	TGx 1848-10E X TGx 1740-2F
18	TGx 1987-9F	Res	TGx 1805-31F X UG5	51	TGx 1951-3F	Susc	TGx1848-10E X TGx 1829-1E
19	UG5	Res	Uganda line	52	TGx 1951-4F	Susc	TGx 1848-10E X TGx 1829-1E
20	TGx 1019-2EB	Susc	TGM 1551 X TGx 536-100C	53	TGx 1954-1F	Susc	TGx 1829-6E X TGx 1448-2E
21	TGx 1019-2EN	Susc	TGm 1552 X TGx 536-100C	54	TGx 1955-4F	Susc	TGx 1838-5E X TGx 1848-10E
22	TGx 1440-1E	Susc	TGx 536-02D X TGx 814-27D	55	TGx 1956-1F	Susc	TGx 1829-6E X TGx 1844-10E
23	TGx 1448-2E	Susc	TGx 824-18D X TGx 814-27D	56	TGx 1961-1F	Susc	TGx 1893-6F X TGX 1805-8F
24	TGx 1485-1D	Susc	TGx 316-024D X TGx813-11D	57	TGx 1965-5F	Susc	TGx 1838-5E X TGx 1848-10E
25	TGx 1740-2F	Susc	TGx 539-5E X SIBLEY	58	TGx 1965-7F	Susc	TGx 1838-5E X TGx 1848-10E
26	TGx 1805-31F	Susc	TGx 1486-1D X TGX 1492-99D	59	TGx 1971-1F	Susc	TGx1893-10F X TGx 1871-12E
27	TGx 1830-20E	Susc	TGx 1483-3D X TGX 1448-1E	60	TGx 1972-1F	Susc	TGx 1895-23F X TGx 1873-16E
28	TGx 1844-18E	Susc	TGx 1449-2D X TGx 1440-1E	61	TGx 1977-2F	Susc	TGx 1805-31F X TGx1864-15F
29	TGx 1844-4E	Susc	TGx 1449-2D X TGx1440-1E	62	TGx 1977-4F	Susc	TGx 1805-31F X TGx 1864-15F
30	TGx 1869-31E	Susc	TGx 1448-2E X TGx 1660-15F	63	TGx 1978-3F	Susc	TGx 1895-4F X TGx 1864-15F
31	TGx 1876-4E	Susc	TGx1660-19F X TGx 1740-6F	64	TGx 536-02D	Susc	TGx 376A X TGM 737
32	TGx 1880-3E	Susc	TGx 1681-3F X TGx 1479-1E	65	TGx 923-2E	Susc	TGx 17-2GE X TGx 849
33	TGx 1893-10F	Susc	TGx 1814-3E X TGx 1740-6F				

[a]Res, resistant; Susc, susceptible.

as total number of different alleles (N_a), number of effective alleles (N_e), Shannon's information index (I), number of private alleles, gene diversity (H_e), observed heterozygosity (H_o) and number of loci with private alleles were computed using Power Marker (Liu and Muse, 2005) and GenAlEx version 6.41 (Peakall and Smouse, 2012) software.

Population structure

The inherent population structure within the genotypes was characterized based on the common attributes of the genotypes using the three complementary clustering approaches. In the first distance-based hierarchical clustering analysis, a pairwise genetic distance (identity-by-state, IBS) matrix was calculated among all individuals using PLINK 5 (Purcell *et al.*, 2007) and Ward's minimum variance. A hierarchical cluster dendrogram was then built from the IBS matrix using the Analyses of Phylogenetics and Evolution (ape) package (Paradis *et al.*, 2004) implemented in R (R core team, 2015). The second approach was a model-based maximum likelihood estimation of ancestral subpopulations using ADMIXTURE (Alexander *et al.*, 2009). ADMIXTURE assumes that the loci are in linkage equilibrium, and the ancestral populations are in Hardy–Weinberg equilibrium (Frichot *et al.*, 2014). In the ADMIXTURE analysis, the number of subpopulations (K) varied from 2 to 12, and the value of K exhibiting a low cross-validation error was selected (Alexander and Lange, 2011). The third approach was an assumption-free discriminant analysis of principal components (DAPC) analysis which was implemented in R using the 'adegenet' package (Jombart, 2008). DAPC that involved optimal clusters of transformed principal component analysis-based SNP data was used to identify and describe clusters of genetically related individuals and subgrouping based on k-means. The best-supported model by Bayesian information criterion (BIC) was selected. Based on results of hierarchical clustering information and ADMIXTURE analysis, the most appropriate K was selected. The membership probabilities of each genotype for the different groups were obtained, and results of the three complementary approaches (the hierarchical tree/dendrogram, ADMIXTURE and DAPC) were compared.

Based on the numbers of inferred clusters determined from the three complementary approaches, analysis of molecular variance (AMOVA) was computed to estimate molecular variation within and among the genotypes using GenAlEx 6.41 and Power Marker V3.25 software. The extent of genetic variance explained by population structure was derived from the AMOVA, fixation index (F_{ST}) and standardized F_{ST} (F'_{ST}) based on Wright's F-statistic (Wright, 1978).

Results

Genetic diversity

The MAF ranged from 0.02 to 0.50, with an average of 0.23. The highest PIC value of 0.38 was recorded in the markers BARC-064873-18956, BARC-051149-11016, BARC-029669-06297 and BARC-019787-04375; while the lowest value was 0.02 with an average value of 0.25. The percentage of polymorphic loci recorded in this study (85%) was high, and an indicator of the efficiency of SNP markers used in this study in detecting polymorphism (Fig. 1). The entire soybean samples had an average effective number of alleles, gene diversity, observed heterozygosity and Shannon's information index of 1.53, 0.31, 0.19 and 0.25, respectively (Table 2). These frequencies were considered to be desirable in differentiating the studied soybean genotypes.

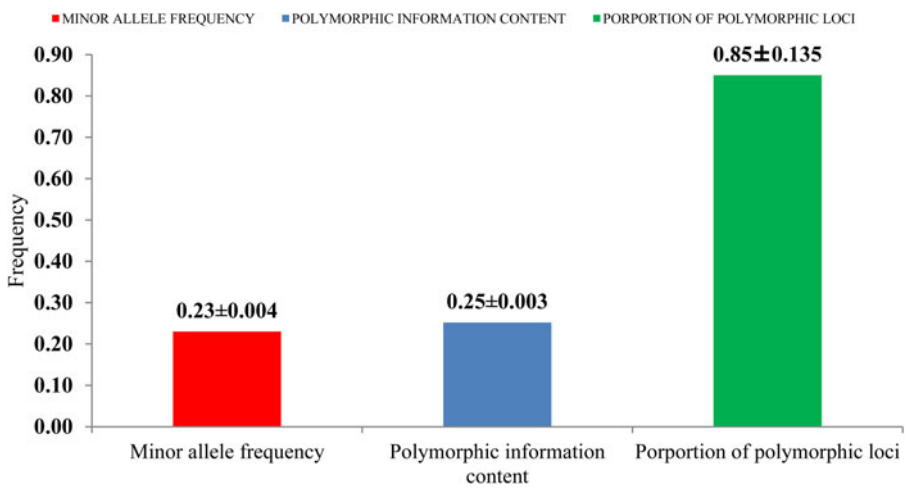

Fig. 1. Frequency distribution of the mean values of genetic properties of SNPs across the 65 soybean genotypes.

Table 2. Average genetic diversity measures and their corresponding standard errors at 1223 SNP loci of the 65 soybean genotypes

	Means			
Diversity parameters	Grand mean	Cluster A	Cluster B	Cluster C
No. of effective alleles	1.533 ± 0.010	1.469 ± 0.011	1.490 ± 0.009	1.455 ± 0.010
Gene diversity	0.314 ± 0.005	0.271 ± 0.005	0.295 ± 0.004	0.273 ± 0.005
Shannon's information index	0.246 ± 0.006	0.407 ± 0.007	0.452 ± 0.006	0.417 ± 0.007
No. of different alleles	1.998 ± 0.008	1.822 ± 0.011	1.971 ± 0.005	1.865 ± 0.010
Observed heterozygosity	0.193	0.048	0.081	0.049
No. private alleles	–	8	78	22

The presence of private alleles was considered as an additional factor to differentiate the population. A total of 108 private alleles were identified among the 65 soybean genotypes. However, based on the number of clusters identified, the number of private alleles detected in cluster 2 was about 3–9-fold higher than the other two clusters.

Genetic relatedness

All the three complementary methods used in determining the number of clusters among the 65 soybean genotypes showed the presence of three major clusters with few sub-clustering (Fig. 2). Cluster 1 consists of 19 genotypes, most of which were crosses with Uganda's UG-5, including the genotype from USA (SOY104), all of which are rust-resistant, except for TGX 536-02D, which was released in Nigeria, Benin and Ghana since 1985, and susceptible to the rust disease. Cluster 2 was the largest, with 26 genotypes, which are all susceptible to leaf rust. Some genotypes (i.e. TGx 1440-1E, TGx 1448-2E, TGx 1937-1F, TGx 1908-8F and TGx 1910-14F) in this cluster are early flowering, resistant to lodging, resistant to pod shattering and have good nodulation ability. These genotypes were released in Nigeria, Cameroon, Togo, Mozambique, Ghana, Benin, Cote D'Ivoire, Kenya, Malawi and Mozambique. Cluster 3 had 20 genotypes, of which only two (TGX 1961-1F and TGX 1835-10E) were resistant to leaf rust with the latter released in Nigeria, Uganda, Kenya and Cameroon since 2008. Other genotypes in this cluster are susceptible to soybean leaf rust disease, but possess other desirable agronomic traits. Although all clusters are discrete and well separated from the other clusters, each cluster is reasonably heterogeneous in terms of the genotypes' attributes. It was observed that genotypes with the same pedigrees clustered together that validates clustering with the SNP markers was efficient in grouping genetically related genotypes. Very few intermixing of the rust-resistant with susceptible soybean genotypes were observed across the clusters (Fig. 2). The DAPC method, using discriminant functions (Fig. 3), maximized the diversity between the three clusters while minimizing the diversity with-in cluster. The first three principal components explained 31.95% of the cumulative variation. The three genetically distinct groups identified using DAPC were consistent with the groups identified by the hierarchical cluster/dendrogram and ADMIXTURE. The error rate from the cross-validation method of both the ADMIXTURE and the BIC from the DAPC showed a rapid decline from $K = 1$ to $K = 3$ and from 1 to 3, respectively, indicating that the samples can be grouped into three major clusters. The results obtained above were consistent and showed good correspondence; thus, indicating the samples' population structure had been correctly identified.

The extent of genetic variation among the genotypes was further revealed by AMOVA, which showed high variations (71%) ($P < 0.001$) among the individual genotypes, and 11% ($P < 0.001$) of the total variation was ascribed to differences among the three clusters detected by the three complementary approaches used to determine the population structure (Table 3). The 18% ($P < 0.001$) was accounted for by the variation within the 65 genotypes. The estimated fixation index (F_{ST}) was 0.11 ($P < 0.001$), indicating moderate genetic differentiation among the clusters.

Discussion

Genetic characterization information is vital in designing future hybridization plans of the IITA soybean breeding programme and partners receiving IITA's elite soybean lines to be utilized by the national soybean improvement programmes across Africa, and beyond. In this study, the discriminatory power and information obtained from the estimates of genetic diversity and population structure of the studied soybean genotypes were enlightening.

The number of alleles per locus measures genetic variation at the gene level. In a population of self-fertilizing species, such as soybean, lower allelic diversity and

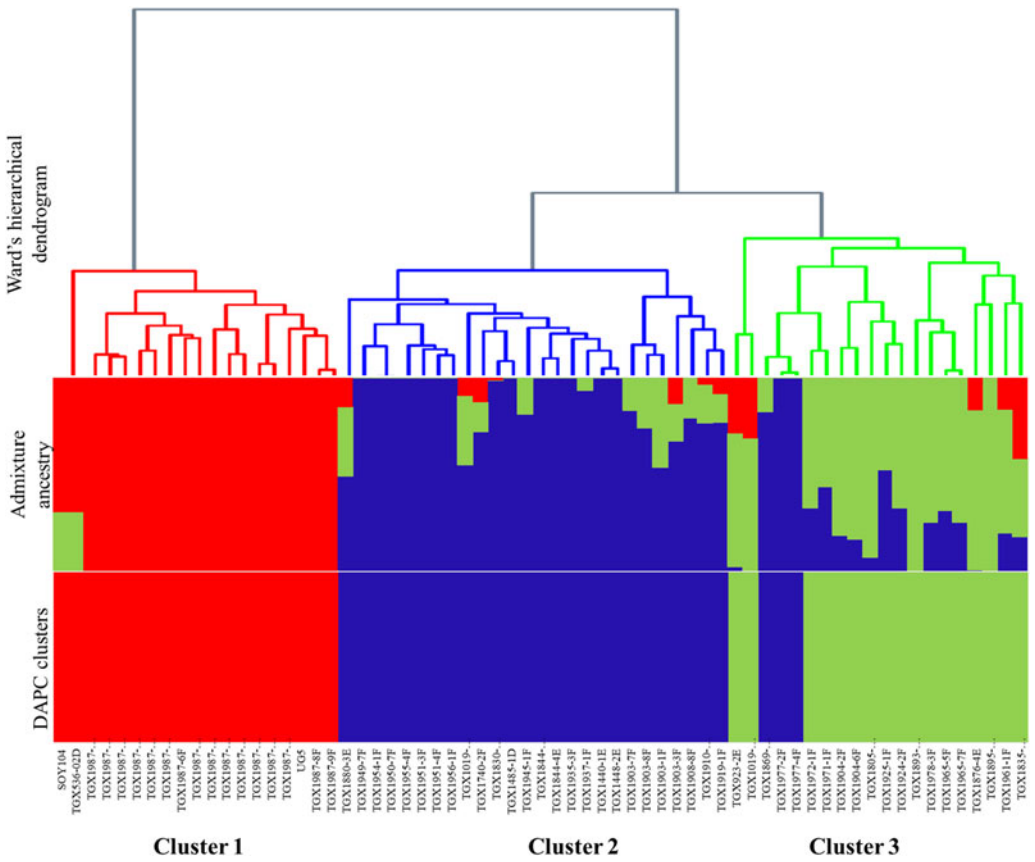

Fig. 2. Hierarchical cluster/dendrogram of the 65 soybean genotypes (Ward's minimum variance method) and the individual ancestry estimated from ADMIXTURE analysis and the clustering by DAPC. Individuals are partitioned into segments corresponding to the inferred membership in $k = 3$ genetic clusters as indicated by the colours.

heterozygosity are commonly expected (Wright, 1921). Hence, PIC, MAF, Shannon's information index, gene diversity and observed heterozygosity were >0.5 in this study. The average PIC value of 0.25 was moderately informative and implies that the SNP markers have differentiating power, since PIC cannot exceed 0.50 in bi-allelic markers (Singh *et al.*, 2013). Comparable average PIC values were reported on ryegrass (Roldàn-Ruiz *et al.*, 2000), soybean (Chen *et al.*, 2017) and wheat (Eltaher *et al.*, 2018). The high percent polymorphic loci and other genetic diversity indices measured also depict the existence of variability among the soybean genotypes. The occurrence of private alleles among the genotypes indicates that the germplasm consisted of diverse, unique, and favourable alleles that may contribute positively to soybean breeding that is yet to be exploited. Thus, these observations imply the presence of diversity within the genotypes and demonstrate that the selected markers were informative and useful for further soybean genetic diversity studies.

For germplasm characterization, earlier reports concluded that large numbers of SNPs would be required to replace the highly polymorphic SSRs in diversity and relatedness studies (Hamblin *et al.*, 2007; Semagn *et al.*, 2014). The average genetic distance (similarity) among a set of genotypes measures genetic diversity at the population level (Lu and Bernardo, 2001). The three diverse but complementary clustering analyses employed in this study differentiated the 65 genotypes from each other, assigning them into three different groups, which indicate substantial genetic diversity. The genotypes in each cluster share common features, which largely corresponds to their pedigree and agronomic traits. TGx 1835-10E, released for leaf rust resistance (Table 1), found in the cluster comprising genotypes with efficient natural nodulation, pod shattering resistance, medium maturity and high yield. The two striga trap varieties: TGx 1440-1E and TGx 1448-2E that are known in depleting the seed bank of *S. hermonthica* through suicidal germination were found in the same cluster with other genotypes that are early maturing, resistant to lodging and high yielding. All the genotypes derived from crosses made to UG-5, a rust-resistant parent, were found in the same cluster because all of them have rust resistance features (Table 1), and hence, according to Tantasawat *et al.* (2011), share genetic homology. Some

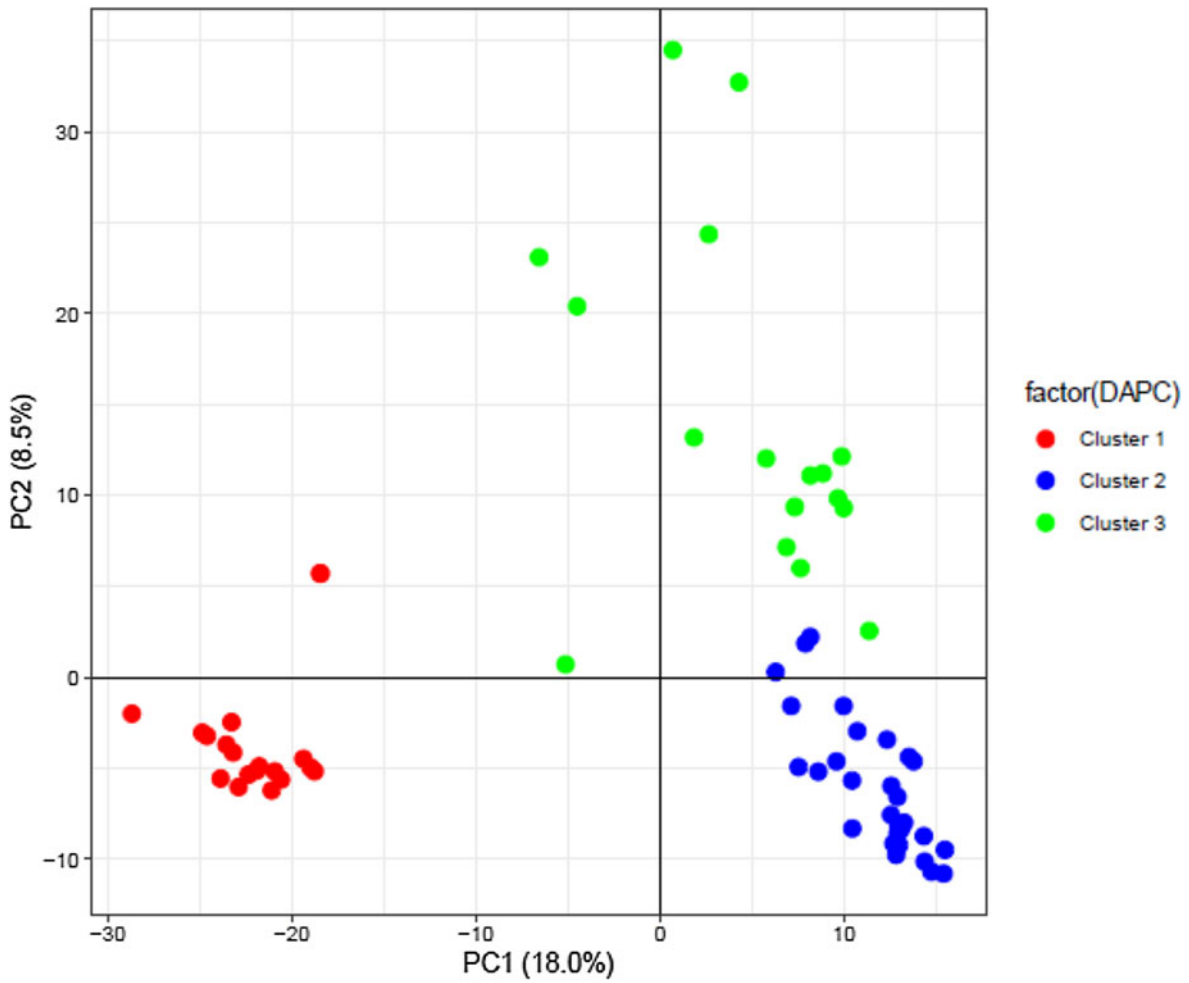

Fig. 3. Scatter plot of individuals on the first two principal components obtained from DAPC based on the analysis of 65 soybean genotypes using 1223 SNP markers. The graph represents the individuals as dots.

correspondence between the clustering pattern and the pedigree of the soybean genotypes was observed. These results indicate that the different pedigrees for each soybean genotypes played an essential role in maintaining genetic variation, as genotypes with similar pedigree clustered together by the SNP markers. Lee *et al.* (2008) suggested that soybean genotypes originating from different genetic backgrounds could have important genetic differences. The markers successfully differentiated the studied germplasm, and the genetic distance observed within the genotypes indicates that good recombination to produce superior progenies can be achieved from crosses between genetically dissimilar genotypes (Narvel *et al.*, 2000).

Assessment of genetic relatedness among parental lines will help breeders identify the most diverse cross

Table 3. Hierarchical AMOVA and Wright's fixation index (F_{ST}) for 65 soybean genotypes based on 1223 SNP markers

Source of variation	df	Sum of squares	Mean squares	Variance components	Percentage of variation	F_{ST}
Among clusters	2	2605.54	1302.77	22.67	11	0.11***
Among genotypes	62	20482.35	330.36	146.52	71	–
Within genotypes	65	2425.50	37.32	37.32	18	–
Total	129	25,513.39	–	206.51	100	–

***Significant at the 0.001 probability level.

combinations useful to enhance genetic gain from the population. Therefore, the relationship of soybean genotypes of IITA breeding lines can facilitate the selection of diverse parental lines carrying priority traits for recombination.

The 71% variation among individual genotypes confirms that populations of self-fertilizing species are expected to have high differentiation. The F_{ST} value in this study was 0.11, which may be regarded as moderate as per Wright's qualitative guidelines (Wright, 1978). These observations indicate the presence of diversity between the genotypes, and demonstrate the highly informativeness and usefulness of the selected markers for future soybean genetic diversity studies.

SNPs have emerged as powerful tools for many genetic applications. Its unique features include abundance in the genome and the ability to generate polymorphism at a single base level. We aim to have the SNP markers optimized by providing relevant information on the markers' discriminatory power in characterizing the IITA soybean germplasm. The markers used were powerful for detecting genetic diversity among and within the soybean populations. We suggest using these validated sets of SNP markers as they spanned the whole genome and provided a biologically sound classification of the genotypes. The results of this study will help to conserve, utilize and manage IITA's soybean germplasm effectively. Determination of the extent of genetic variability present in this germplasm will furnish soybean breeders' with the required information and decision-making tools for effective parental selection in their breeding programmes.

Acknowledgements

This research was supported by the core fund provided by the International Institute of Tropical Agriculture (IITA), and support from Soybean Genomics and Improvement Laboratory, Beltsville Agricultural Research Center (SGIL-BARC), Beltsville, Maryland.

Conflict of interest

None.

References

Alexander DH and Lange K (2011) Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* 12: 246.

Alexander DH, Novembre J and Lange K (2009) Fast model based estimation of ancestry in unrelated individuals. *Genome Research* 19: 1655–1664.

Chen Y and Nelson RL (2005) Relationship between origin and genetic diversity in Chinese soybean germplasm. *Crop Science* 45: 1645–1652.

Chen W, Hou L, Zhang Z, Pang X and Li Y (2017) Genetic diversity, population structure, and linkage disequilibrium of a core collection of *Ziziphus jujuba* assessed with genome-wide SNPs developed by genotyping by-sequencing and SSR markers. *Frontiers in Plant Science* 8: 575.

Chigeza G, Boahen S, Gedil M, Agoyi E, Mushoriwa H, Denwar N, Gondwe T, Tesfaye A, Kamara A, Alamu OE and Chikoye D (2019) Public sector soybean (*Glycine max*) breeding: advances in cultivar development in the African tropics. *Plant Breeding* 2019: 1–10.

Cornelious BK and Sneller CH (2002) Yield and molecular diversity of soybean lines derived from crosses of northern and southern elite parents. *Crop Science* 42: 642–647.

Dellaporta SL, Wood J and Hicks JB (1983) A plant DNA mini preparation: version II. *Plant Molecular Biology Reporter* 1: 19–21.

Eltaher S, Sallam A, Belamkar V, Emara HA, Nower AA, Salem KF, Poland J and Baenziger PS (2018) Genetic diversity and population structure of F3:6 Nebraska winter wheat genotypes using genotyping-by-sequencing. *Frontiers in genetics* 9: 76.

FAO (2020) FAOSTAT. URL: http://www.fao.org/faostat/en/#data. Accessed date and time: 12/10/202, 9:40 am.

Frichot E, Mathieu F, Trouillon T, Bouchard G and François O (2014) Fast and efficient estimation of individual ancestry coefficients. *Genetics* 196: 973–983.

Gizlice Z, Carter Jr TE and Burton JW (1993) Genetic diversity in North American soybean: I. Multivariate analysis of founding stock and relation to coefficient of parentage. *Crop Science* 33: 614–620.

Gizlice Z, Carter Jr TE and Burton JW (1994) Genetic base for North American public soybean cultivars released between 1947 and 1988. *Crop Science* 34: 1143–1151.

Gizlice Z, Carter Jr TE, Gerig TM and Burton JW (1996) Genetic diversity patterns in North American public soybean cultivars based on coefficient of parentage. *Crop Science* 36: 753– 765.

Hamblin MT, Warburton ML and Buckler ES (2007) Empirical comparison of simple sequence repeats and single nucleotide polymorphisms in assessment of maize diversity and relatedness. *PLoS One* 2: 1367.

Helms T, Orf J, Vallad G and McClean P (1997) Genetic variance, coefficient of parentage, and genetic distance of six soybean populations. *Theoretical and Applied Genetics* 94: 20–26.

Hymowitz T and Shurtleff WR (2005) Debunking soybean myths and legends in the historical and popular literature. *Crop Science* 45: 473–476.

Hyten DL, Song Q, Choi IY, Yoon MS, Specht JE, Matukumalli LK, Nelson RL, Shoemaker RC, Young ND and Cregan PB (2008) High-throughput genotyping with the GoldenGate assay in the complex genome of soybean. *Theoretical and Applied Genetics* 116: 945–952.

Jombart T (2008) Adegenet an R package for the multivariate analysis of genetic markers. *Bioinformatics (Oxford, England)* 24: 1403–1405.

Keim P, Beavis GW, Schupp J and Freestone R (1992) Evaluation of soybean RFLP marker diversity in adapted germplasm. *Theoretical and Applied Genetics* 85: 205–212.

Kisha T, Sneller CH and Diers BW (1997) Relationship of genetic distance and genetic variance in populations of soybean. *Crop Science* 37: 1317–1325.

Kisha T, Diers BW, Hoyt JM and Sneller CH (1998) Genetic diversity among soybean plant introductions and North American germplasm. *Crop Science* 38: 1669–1680.

Lee JD, Yu JK, Hwang YH, Blake S, So YS, Lee GJ, Nguyen HT and Shannon JG (2008) Genetic diversity of wild soybean

(*Glycine soja* Sieb. and Zucc.) accessions from South Korea and other countries. *Crop Science* 48: 606–616.

Liu K and Muse SV (2005) Power marker: an integrated analysis environment for genetic marker analysis. *Bioinformatics (Oxford, England)* 21: 2128–2129.

Lu H and Bernardo R (2001) Molecular marker diversity among current and historical maize inbreds. *Theoretical and Applied Genetics* 103: 613–617.

Narvel JM, Fehr WR, Chu WC and Grant D (2000) Simple sequence repeat diversity among soybean plant introductions and elite genotypes. *Crop Science* 40: 1452–1458.

Nelson LA, Elmore RW, Klein RN and Shapiro C (1998) *Nebraska Soybean Variety Tests. Nebraska Coop. Ext. E. C. 98-104-A*. Lincoln: University of Nebraska.

Paradis E, Claude J and Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics (Oxford, England)* 20: 289–290.

Peakall R and Smouse P (2012) Genalex 6.5: genetic analysis in Excel. Population genetic software for teaching and research – an update. *Bioinformatics (Oxford, England)* 28: 2537–2539.

Purcell S, Neale B, Todd-brown K, Thomas L, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ and Sham PC (2007) PLINK: a toolset for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* 81: 559–575.

Qiu L, Chen P, Liu Z, Li Y, Guan R, Wang L and Chang R (2011) The worldwide utilization of the Chinese soybean germplasm collection. *Plant Genetic Resources: Characterization and Utilization* 9: 109–122.

R Core Team (2015) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Roldàn-Ruiz I, Dendauw J, Van Bockstaele E, Depicker A and De Loose MA (2000) AFLP markers reveal high polymorphic rates in ryegrasses (*Lolium* spp.). *Molecular breeding* 6: 125–134.

Salado-Navarro LR, Sinclair TR and Hinson K (1993) Changes in yield and seed growth traits in soybean cultivars released in the southern U.S.A. from 1945 to 1983. *Crop Science* 33: 1204–1209.

Semagn K, Babu R, Hearne S and Olsen M (2014) Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): overview of the technology and its application in crop improvement. *Molecular Breeding* 33: 1–14.

Singh RJ and Hymowitz T (1999) Soybean genetic resources and crop improvement. *Genome* 42: 605–616.

Singh N, Choudhury DR, Singh AK, Kumar S, Srinivasan K, Tyagi RK, Singh NK and Singh R (2013) Comparison of SSR and SNP markers in estimation of genetic diversity and population structure of Indian rice varieties. *PLoS One* 8: e84136.

Sneller CH (1994) Pedigree analysis of elite soybean lines. *Crop Science* 34: 1515–1522.

Sneller CH, Miles J and Hoyt JM (1997) Agronomic performance of soybean plant introduction and their genetic similarity to elite lines. *Crop Science* 37: 1595–1600.

Soybase (2020) Soybean genetic resources and genetic enhancements white paper. URL: https://soybase.org/Genetic_Resources/Soybean_Genetic_Resources.html accessed date: 16/08/2020.

Tantasawat P, Trongchuen J, Prajongjai T, Jenweerawat S and Chaowiset W (2011) SSR analysis of soybean (*Glycine max* (L.) Merr.) genetic relationship and variety identification in Thailand. *Australian Journal of Crop Science* 5: 283–290.

Tefera H, Kamara AY and Asafo-Adjei B (2009) Improvement in grain and fodder yields of early maturing promiscuous soybean varieties in the Guinea savanna of Nigeria. *Crop Science* 49: 2037–2042.

Wright S (1921) Systems of mating. II. The effects of inbreeding on the genetic composition of a population. *Genetics* 6: 124–143.

Wright S (1978) *Evolution and the Genetics of Populations Vol. 4. Variability Within and among Natural Populations*, Chicago: University of Chicago Press, p. 58.