Genetic Diversity and Population Structure Analysis of Tropical Soybean (*Glycine Max* (L.) Merrill) using single Nucleotide Polymorphic Markers

By Tonny Obua, Julius P. Sserumaga, Stephen O. Opiyo, Phinehas Tukamuhabwa, Thomas L. Odong, Josiah Mutuku & Nasser Yao

*Makerere University Kampala*

**Abstract**—Soybean (*Glycine max* (L.) Merrill) is among the most important crops worldwide due to its numerous uses in feed, food, biofuel, and significant atmospheric nitrogen fixation capability. To understand the genetic diversity and population structure of tropical soybean germplasm, 89 genotypes from diverse sources were analyzed using 7,962 SNP markers. The AMOVA results showed low diversity among and high within the populations, while the polymorphism information content (PIC) was 0.27. Both phylogenetic and principal component analysis grouped the 89 soybean genotypes into three major clusters, while population structure grouped the soybean genotypes into two subpopulations. On the other, the average Roger genetic distances within the study population was 0.34. The low diversity reported in the studied soybean germplasm pool is particularly worrying, considering the new trends of climate change and the emergence of new pests and diseases of soybean. Therefore, in order to address these challenges and develop soybean varieties with desirable traits, there is a need to broaden the genetic base of tropical soybean through the importation of germplasm from other countries.

**Keywords:** genetic diversity, population structure, single nucleotide polymorphism (SNP), soybean, tropical soybean genotypes.

**GJSFR-D Classification:** FOR Code: 070199

Strictly as per the compliance and regulations of:

© 2020. Tonny Obua, Julius P. Sserumaga, Stephen O. Opiyo, Phinehas Tukamuhabwa, Thomas L. Odong, Josiah Mutuku & Nasser Yao. This is a research/review paper, distributed under the terms of the Creative Commons Attribution 4.0 Generic License [https://creativecommons.org/licenses/by/4.0/](https://creativecommons.org/licenses/by/4.0/), permitting all commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
Genetic Diversity and Population Structure Analysis of Tropical Soybean (Glycine Max (L.) Merrill) using single Nucleotide Polymorphic Markers

Tonny Obua ⁺, Julius P. Sserumaga ⁺, Stephen O. Opiyo ⁺, Phinehas Tukamuhabwa ⁺, Thomas L. Odong *, Josiah Mutuku § & Nasser Yao *

Abstract: Soybean (Glycine max (L.) Merrill) is among the most important crops worldwide due to its numerous uses in feed, food, biofuel, and significant atmospheric nitrogen fixation capability. To understand the genetic diversity and population structure of tropical soybean germplasm, 89 genotypes from diverse sources were analyzed using 7,962 SNP markers. The AMOVA results showed low diversity among and high within the populations, while the polymorphism information content (PIC) was 0.27. Both phylogenetic and principal component analysis grouped the 89 soybean genotypes into three major clusters, while population structure grouped the soybean genotypes into two subpopulations. On the other, the average Rogers genetic distances within the study population was 0.34. The low diversity reported in the studied soybean germplasm pool is particularly worrying, considering the new trends of climate change and the emergence of new pests and diseases of soybean. Therefore, in order to address these challenges and develop soybean varieties with desirable traits, there is a need to broaden the genetic base of tropical soybean through the importation of germplasm from other countries.

Keywords: genetic diversity, population structure, single nucleotide polymorphism (SNP), soybean, tropical soybean genotypes.

I. Introduction

Soybean is among the most valuable crops in Uganda and across the East African region due to the high protein content that makes it an important ingredient in the diets of the people and livestock (Tukamuhabwa, 2001). Several soybean processing plants have been established in Uganda and across the East African region with large processing capacities to develop different products from soybean. This new development has motivated the farmers to produce more grains to supply these plants (Tukamuhabwa et al., 2019). The three leading African countries in soybean production are South Africa (1,540,000 MT), Nigeria (758,033 MT), and Zambia (302,720 MT) (FAO 2018). Uganda is 11th in Africa and 1st in East Africa, with a production of 29,000 MT (FAO 2018). Hence soybean production and consumption have led to increased farmers’ income, improved food and nutrition security, and poverty eradication at the rural household level (Ssengendo et al., 2010; SNV, 2011; Tukamuhabwa & Obua, 2015). Accordingly, soybean has the potential to contribute to poverty alleviation in Uganda and across the East African region.

Despite the contribution of soybean to smallholder farmers in Uganda and across the East African region, development of new varieties has been hindered by the low genetic diversity of the crop that have been observed in other countries (Gupta & Manjaya, 2017; Kumawat et al., 2015; Liu et al., 2017; Maldonado dos Santos et al., 2016; Torres et al., 2015). Kumawat et al. (2015) investigated the diversity of 82 Indian soybean accessions using SSR markers and identified three major clusters. In another study, Torres et al. (2015) found that both Principal Component Analysis (PCA) and STRUCTURE, clustered 191 soybean accessions in Brazil into two groups. Similarly, Gwinner et al. (2017) in another study that aimed at understanding the genetic diversity and population structure of 77 commercial soybean varieties in Brazil using 35 SSR markers, reported low genetic diversity in soybean germplasm.

To assess the genetic diversity of soybean and other plants, various methods such as morphological markers, geographic origins, pedigree information, isozymes, and DNA markers have been applied (Dayaman, 2007; Appiah-Kubi, 2012; Ojo et al., 2012; Malek et al., 2014; Villela et al., 2014). The use of morphological trait has remained a powerful taxonomic tool for preliminary grouping of germplasm before their classification using more precise marker techniques.
In fact, several studies involving the classification of plants still rely on the use of morphological traits (Khalid et al., 2010). Additionally, the use of morphological markers in classification is easy to score, cheaper and fast. However, the disadvantage of using morphological markers is that it’s less robust compared to most molecular markers and outcomes can be easily influenced by environmental factors. In the case of pedigree information, limitations such as uncertain and incomplete data errors are likely, while for isozymes, chances of limited data are more prominent (Li & Nelson, 2001; Wang et al., 2006). So far, DNA markers remain the most precise method of genetic diversity analysis that have been complemented with morphological trait analysis. Among different DNA markers, random amplified polymorphic DNAs (RAPDs), amplified fragment length polymorphisms (AFLPs), restriction fragment length polymorphisms (RFLPs), simple sequence repeats (SSRs) and single nucleotide polymorphisms (SNPs) have been widely used in understanding the diversity of soybean; each with its advantages and disadvantages (Chauhan et al., 2015; Chen et al., 2017; Doldi et al., 1997; Hipparagi et al., 2017; Ojo et al., 2012; Ren et al., 2013; Singh et al., 2013; Tantasawat et al., 2011; Tanya et al., 2001; Torres et al., 2015). SSR markers have been widely used to determine genetic diversity in many crops because they are easy to use, reasonably low price, and high level of polymorphism (Vignal et al., 2002). However, recently SNP markers have been widely utilized for assessment of diversity in plants because they occur much more frequently in the genome than SSR markers, and their genotyping can be easily automated (Mammadov et al., 2012). In the current study, we used Genotype By Sequencing (GBS) technology to study a collection of 89 tropical soybean germplasm collected from different countries. Therefore, the objective of the current study was to understand the genetic diversity and population structure of tropical soybean germplasm using SNP markers. Since the genotypes included in the current study are parental lines, land races, released varieties, and advanced lines, they are representative of the existing germplasm in tropical Africa.

II. Materials and Methods

a) Plant materials

In this study, we used a total of 89 tropical soybean genotypes; these included collections from different sources that possess high genetic diversity (45 genotypes were from Uganda, 13 from Japan, six from the USA, 12 from World Vegetable Center (AVRDC) in Taiwan and 13 from Seed Co; a seed Company from Zimbabwe (Supplementary table).
calculated for each value of K using Structure Harvester (Earl, Cruz, and Vonholdt 2012; Evanno et al. 2005). A line was assigned to a given cluster when the proportion of its genome in the cluster (qK) was higher than a standard threshold value of 70%. For the chosen optima value of K, membership coefficient matrices of replicates from STRUCTURE were integrated to generate a Q matrix using the software CLUMPP (Jakobsson and Rosenberg, 2007) and the STRUCTURE bar plot was drawn using the DISTRACT software (Rosenberg, 2004). Principal coordinate analysis was performed based on the genetic distance matrix using the Dissimilarity Analysis and Representation for windows (DARwin) v.6.0.013 (http://darwin.cirad.fr). To validate and gain more insight into the genetic diversity of the soybean germplasm panel used in this study, we generated a phylogenetic tree by the neighbor-joining method. Analysis of molecular variance (AMOVA) was performed using GenALEx V6.5 software.

III. Results

a) Genotype Diversity analysis

A total of 16,688 SNP markers were identified in the 89 genotypes of soybean; of those 7,962 polymorphic and non-redundant SNP markers, with greater than 5% minor allele frequency (MAF) and missing data lower than 20% were used for subsequent analysis. These 7,962 SNPs detected a total of 15,924 alleles as expected. The average PIC was 0.27, ranging from 0.01 to 0.50, and heterozygosity ranged from 0.0 to 0.35 of individuals and 0.0 to 0.8 of markers (Fig. 1).

![Fig. 1: Levels of heterozygosity of individual soybean genotypes and SNPs markers](image)

b) Genetic distance and relationship

The average Roger genetic distances within the study population was 0.34. From a total of 89 genotypes, 18.1% of the distance values were between 0.0 and 0.05, while 20.7% were between 0.35 and 0.40 (Fig. 2). Relative kinship reflects the approximate degree of identity between two given genotypes. For combined analysis of all 89 genotypes, the kinship coefficients ranged from 0 to 1.04, with an overall average of 0.51; only 1.6% of the pair wise kinship estimates had values of 0.0 – 0.05 while 76.1% had values ranging from 0.5 – 0.550, indicating that most of the genotypes were in one way or another related and very few genotypes were not related (Fig. 3).
Fig. 2: Distribution of pairwise Roger’s genetic distance calculated for 89 soybean genotypes

Fig. 3: Distribution of pair-wise kinship coefficients among 89 soybean genotypes from different sources based on 7,692SNPs

c) Population structure analysis

The log probability of the data LnP (D) increased continuously with increasing K (number of groups or populations). The ad hoc statistic ΔK showed a higher likelihood value at K = 2 as the highest level of structure (Fig. 4). This pattern was also observed in the population structure, where two groups were formed (Fig. 5).
Fig. 4: Estimation of population using LnP(D) derived ΔK with K ranged from 1 to 10 with 7,692 SNPs

Fig. 5: Population structure (K = 2) inferred from STRUCTURE analysis for the 89 soybean genotypes based on 7,692 SNPs

d) Neighbor-joining Phylogenetic Tree

The phylogenetic tree grouped the 89 soybean genotypes into three major clusters (Fig 6). The genotypes were separated into three distinct sub-clusters: There were 40 genotypes in sub-cluster 1, which included Nam II and GC00138-29, and 13 progenies derived from a cross between these two genotypes. Nam II is a Ugandan variety, which is a selection from TGM 79; obtained from IITA while GC00138-29 is a variety from AVRDC in Taiwan. This sub-cluster also included released varieties in Uganda; Namsoy 3 which is a cross between Kabanyolo 1 and Nam 1 (selection from ICAL 131 from the USA), and Maksoy 5N that is a progeny of Nam II and GC00138-29. The second sub-cluster had 26 genotypes, among which 13 genotypes were from Seed co in southern Africa and eight genotypes from AVRDC, Taiwan. It was surprising that Namsoy 4M, a released Ugandan variety that is a progeny of Nam II and GC00138-29, was clustered in this sub-cluster. By comparison, the other remaining 23 genotypes belonged to sub-cluster 3, among which seven genotypes were progenies from a cross between Duiker and TGx 1835-10E while nine were from a cross between Duiker and GC00138-29. This sub-cluster also included released Ugandan varieties, Maksoy 1N (selection from TGx 1835-10E), Maksoy 2N (Duiker X TGx 1835-10E), Maksoy 3N and Maksoy 4N (Duiker X GC00138-29). However, few S lines and AVRDC genotypes were scattered in all three major clusters.
e) **Principal Component Analysis (PCA)**

PCA has been suggested as an alternative to population structure analysis for studying population stratification from genotypic data (Patterson et al., 2006). A PCA of the 89 genotypes with the 7,962 SNPs also showed a clear separation of the same three major groups that were identified by the phylogenetic tree (Fig. 7).

![Fig. 6: Tree based on the Neighbor Joining method showing genetic dissimilarity between soybean genotypes, based on SNP markers](image)

**Fig. 6:** Tree based on the Neighbor Joining method showing genetic dissimilarity between soybean genotypes, based on SNP markers

![Fig. 7: Plot of PC1 (40.6%) and PC2 (18.2%) from principal coordinate analysis based on genetic distance matrix calculated for 89 soybean genotypes genotyped with 7,692SNPs](image)

**Fig. 7:** Plot of PC1 (40.6%) and PC2 (18.2%) from principal coordinate analysis based on genetic distance matrix calculated for 89 soybean genotypes genotyped with 7,692SNPs

f) **Analysis of molecular variance**

Analysis of molecular variance (AMOVA) among the 89 soybean genotypes indicated that 2% of the variance was due to genetic differentiation among the populations, 98% of the variance was accounted for by genetic differentiation among individuals within populations.

IV. **Discussion**

One of the requirements for a successful breeding program is a high level of genetic diversity among the germplasm used for the development of new crop varieties. Over the years, most soybean breeding programs have replaced traditional varieties or
landraces with more modern varieties with desirable attributes that have led to increased yields. However, in the current study, to compare the genetic diversity of tropical soybean genotypes, we studied fairly diverse sets of genotypes from Uganda, Zimbabwe, Japan, Taiwan, and the USA. These genotypes included parental lines, land races, released varieties, and advanced lines that are representative of the existing germplasm in tropical Africa.

The level of genetic diversity observed in this study is lower compared to previously reported results based on SNP data (Li et al. 2010; Hao et al. 2012; Zhou et al. 2015). The observed low diversity is because the genotypes used in the present study were mainly released varieties and advanced breeding lines. In contrast, the genotypes used in Li et al. (2010), Hao et al. (2012) and Zhou et al. (2015) included mainly wild relatives and landraces of soybean. On the other hand, previous studies that involved improved soybean varieties also observed low genetic diversity (Liu et al. 2017; Maldonado dos Santos et al. 2016). These improved varieties tend to have low genetic diversity because of the high selection pressure subjected to the genotypes during evaluation and selection (Gwinner et al. 2017). This was also confirmed by genetic distance and kinship analysis that showed that majority of the genotypes in this study are related to each other in one way or another.

The phylogenetic tree and PCA analyses indicated the existence of three major sub-clusters among the 89 genotypes of our study. On the other hand, population structure clustered the genotypes into two major subpopulations. Sub-cluster 1 included Nam II and GC00138-29 and 13 progenies derived from a cross between these two genotypes. Nam II is a Ugandan variety, which is a selection from TGM 79; obtained from IITA while GC00138-29 is a variety from AVRDC in Taiwan. This sub-population also included Maksoy 5N, released in 2013 and NII X GC 44.2 that was released in 2017 as Maksoy 6N and are progenies of Nam II and GC00138-29 cross. Since TMG 79 and GC00138-29 were introduced to Uganda through Consultative Group for International Agricultural Research (CGIAR) institutions that usually collect germplasm from different countries, there is a possibility that they share the same geographical origin. On the other hand, genotypes from Seed Co and AVRDC, Taiwan were grouped in the second sub-cluster. This implies that soybean varieties from Seed Co share very similar parents and geographical origin with genotypes from AVRDC.

By comparison, the third sub-cluster mainly consisted of progenies from two crosses; Duiker X TGx 1835-10E and Duiker X GC00138-29. The sub-cluster also included released Ugandan varieties; Maksoy 1N (selection from TGx 1835-10E), Maksoy 2N (Duiker X TGx 1835-10E), Maksoy 3N and Maksoy 4N (Duiker X GC00138-29). Duiker originated from Zimbabwe and was used as a female parent during generation of the two crosses.

V. Conclusion

Genetic variation and population structure of the core germplasm available for soybean breeding in Uganda and the East African region were assessed using high-density SNP markers. The results of the study showed a low level of heterogeneity within most of the genotypes studied, suggesting that the current generation of inbreeding has fixed lines. The observed low diversity in the germplasm pool is particularly worrying; considering the vulnerability of agriculture under the impact of climate changes. For example, in Uganda, we have observed the emergence of two new soybean pests (groundnut leaf miner and bruchids) that previously were not main production constraints. This is coupled with breakdown of soybean rust resistance in the existing soybean varieties in Uganda that were previously resistant to the disease due to several virulent races of soybean rust pathogen. Therefore addressing these challenges and developing soybean varieties with the desirable traits, requires diversification of the genetic background of the current breeding population by incorporating new genetic resources from other countries.

Acknowledgments

This project was supported by the BecA-ILRI Hub through the Africa Biosciences Challenge Fund (ABCF) program. The ABCF Program is funded by the Australian Department for Foreign Affairs and Trade (DFAT) through the BecA-CSIRO partnership; the Syngenta Foundation for Sustainable Agriculture (SFSA); the Bill & Melinda Gates Foundation (BMGF); the UK Department for International Development (DFID) and; the Swedish International Development Cooperation Agency (Sida).

References Références Referencias


Markers and Their Impact on Plant Breeding
International Journal of Plant Genomics Article ID 728398.


Compliance with Ethical Standards
Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors
Conflict of interest: The authors declare no conflict of interest.