Genetic diversity and population structure of Ugandan soybean (*Glycine max* (L.) germplasm based on DArTseq

Musondolya Mathe Lukanda (✉ lukandamathe6@gmail.com)
Makerere University

Isaac Onziga Dramadri
Makerere University

Emmanuel Amponsah Adjei
Makerere University

Arusei Perpetua
Makerere University

Hellen Wairimu Gitonga
Makerere University

Peter Wasswa
Makerere University

Richard Edema
Makerere University

Mildred Ochwo Ssemakula
Makerere University

Phinehas Tukamuhabwa
Makerere University

Tusiime Geoffrey
Makerere University

Research Article

Keywords: Diversity, STRUCTURE, Polymorphism, Clustering, Soybean, DArTseq

Posted Date: August 1st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1689218/v2

Abstract

Soybean (*Glycine max* (L.) is an important oil crop with production below the real need in Africa. To increase the production, substantial upgrading must be accomplished by enlarging the genetic potential of new cultivars that relays on the parents' genetic diversity. We aimed to assess the genetic diversity and the population structure of soybean accessions, therefore, evaluate soybean value in terms of use for breeding purposes. To accomplish this, a set of 282 soybean accessions were genotyped using the Diversity Array Technology Sequencing with a high throughput of the Single-nucleotide polymorphisms (SNPs). A total of 6,935 high-quality SNPs were identified across the whole soybean genome. The mean value of genetic diversity, major alleles frequency, minor alleles frequency, expected heterozygosity, and the polymorphism information content was respectively estimated at 0.35, 0.77, 0.22, 0.33, and 0.29. The analysis of molecular variance reveals that the variance among the populations is lower than the variance within the populations. The soybean accessions could be classified into two groups based on the model-based population structure and the principal coordinate analysis or into three groups based on the discriminant analysis of the principal component and the Neighbor-joining tree. The information provided in this study will be helpful for breeders.

Introduction

Soybean (*Glycine max* (L.) is classified as a miracle crop because it serves as a source of oil (Khojely et al., 2018), protein (Khojely et al., 2018; Naik et al., 2016) and also provides biofuel for industrial use (Mutegi and Zingore, 2013; Naik et al., 2016). According to Hartman et al., (2011), soybean is the fourth most widely grown crop worldwide. It accounts for close half of the total production of oilseeds (Chander et al., 2019; Foster et al., 2009) and in economic development, it contributes to farmers' income, food security, animal feed, and industry development (Hartman et al., 2011). Among oilseed crops, soybean has a low relatively oil content (Zimmer, 2010) but the soybean cropping is adapted to different climatic and soil conditions (Sinclair et al., 2014). In addition, the capacity of soybean to fix nitrogen contributes to improving soil fertility, which offers the soybean a competitive choice in crop rotation in Africa at the household level (Obua et al., 2020; Tefera et al., 2009), and its expansion in Africa cropping system (Khojely et al., 2018).

Studies have proven that Soybean varieties with appreciable yield performance are needed to sustain the production but their improvement has been slowed down by the low genetic diversity (Clever et al., 2020; Jo et al., 2021). Soybean is a predominantly self-pollinate crop (Chiari et al., 2005; Zhao et al., 2018) and has been in relating more to its genetic diversity. To tackle production challenges and value chain stakeholders' there is the need to broaden the diversity of the raw breeding material for breeding purposes (Carter et al., 2016).

Genetic diversity is among indicators that breeding programs target to enlarge the potentiality of a set of populations used for developing new varieties (Fu, 2015; Govindaraj et al., 2015). Understanding genetic diversity is a process by which variation among groups of individuals or individuals or populations is analyzed using a specific method or a combination of methods (Mohammadi and Prasanna, 2003; Mukhopadhyay and Bhattacharjee S., 2016). Recently, several studies applied those tools in soybean genetic diversity and population structure analysis (Fried et al., 2018; Gupta, 2017; Nkongolo et al., 2020a; Obua et al., 2020).

Morphological and biochemical markers are affected by the environment hence possesses a lot of limitations for breeding purposes however, molecular markers are considered as a heritable polymorphism that can be measured in a group of individuals or a single individual (Gudeta, 2018; Raza et al., 2016) and are non-sensitive to the environment. Therefore, molecular markers are robust and invaluable tools for genomic analysis (Duran et al., 2009; Raza et al., 2016). The molecular markers offer an advantage in their rapidity and freedom growth stage-specificity (Chander et al., 2021). Furthermore, molecular markers techniques reduced the required amounts of tissue samples needed, thus facilitating the analysis of single seeds or seedlings (Nadeem et al., 2018). Thus, a considerable number of molecular markers have been used in soybean genetic diversity analysis. Those studies reported a narrow genetic diversity (Jo et al., 2021), or a moderate genetic divisity (Gupta SK, 2017; Nkongolo et al., 2020), or a high genetic diversity (Kujane et al., 2019; Torres et al., 2015) in soybean genotypes. They recommended the findings to the soybean breeders in selecting genetically distinct parents for a soybean improvement. Uganda is one of the African soybean leader producers and has established a soybean breeding program with considerable germplasm that aims to respond to the increasing demand for soybean processors in Uganda and East African regions (Khojely et al., 2018; Tukamuhabwa et al., 2016). Understanding the genetic diversity and population structure of soybean accessions serves as a gateway to meeting the community'requirement for new cultivars development. Previous studies by Clever et al., (2022) and Obua et al., on soybean genetic diversity reported less diversity in the raw breeding soybean materials. Hence, they recommended enlarging the genetic diversity of soybean collection to create a genetic variation that is necessary to cope with the dynamics of biotic and abiotic stresses that affect soybean production in Uganda (Clever et al., 2020; Obua et al., 2020). But in those studies, few accessions were involved and figuring the availability or importation of new soybean accessions as part of the solution, this study sought to the genetic diversity and population structure of a set of 282 soybean lines from Zimbabwe, Taiwan, the USA, and Uganda using the Diversity Array Technology Sequencing (DArT-Seq).

Materials And Methods

Plant materials

Two hundred eighty-two (282) soybean accessions were collected from the National Crops Resources Research Institute (NaCRRI) at Namulonge-Uganda and the Makerere University which sourced from four different countries including Uganda, Taiwan, the USA, and Zimbabwe (S1 Table).

DNA extraction and soybean genotyping

Seeds from a set of 282 soybean accessions were planted at the Makerere University Agricultural Research Institute Kabanyoro (MUARIK). Fifteen days after, the fresh leaves were collected and kept on three 96-well plates. The three plates were expedited to the Integrated Genotyping Service and Support of the Biosciences in Eastern and Central Africa—ILRI Hub, Kenya, for genotyping. The DNA was extracted from the leaf tissues using the Nucleomag Plant Genomic DNA extraction kit (Macherey-nagel GmbH, 2018), and the DNA quality check was conducted on 0.8% agarose. Genotyping was performed using Diversity Array Technology sequencing (DArTseq). Then, a genomic DNA library was constructed using genomic complexity reduction technology (Kilian et al., 2012). The library was purified and quantified for cluster generation in an automated clonal amplification system (cBOT Illumina). Thereafter, next-generation sequencing was performed using the sequencer HiSeq 2500 (Illumina).

Data analysis

The Data quality control, the filtering, and the imputation were performed using the TASSEL Software v.5.73. Single Nucleotide Polymorphism (SNP) markers with more than 20% of missing data, a minor allele frequency (MAF) of 0.05, or an unfixed position were removed (Adu et al., 2019; Chander et al., 2021). The imputation was performed using the LD KNNI algorithm, based on a *k*-nearest neighbor genotype imputation method (Money et al., 2015; Troyanskaya et al., 2001). After the filtering, four soybean lines were excluded from the original set of 282 lines due to their low sample quality control and high missing data (≥20% missing information) rate. Hence, we proceeded with the analysis with a set of 278 soybean genotypes. The summary statistics including major and minor alleles frequencies were generated using TASSEL Software v.5.73 whilst computer statistics such as the gene diversity and the polymorphism information content (PIC) using Power Marker v3.25 (K. Liu & Muse, 2005). Observed and expected heterozygosity was generated using the “Adegenet” package in R Software (Jombart et al., 2021).

After the imputation, the SNPs data was submitted to the genetic analysis of the population structure through a Bayesian clustering approach in STRUCTURE v.2.3.4 (Evanno et al., 2005; Porras-Hurtado et al., 2013). The structure analysis was run considering a burn-in period of 10, 000 Markov-chain Monte Carlo iterations and a 10,000-run length with an admixture model following the Hardy-Weinberg equilibrium and its correlated allele frequencies (Chander et al., 2021). Ranged values from 1 to 10 of the number of clusters (K) were performed and run independently. The structure outputs were analyzed using Structure Harvester which enabled the identification of the best K value as the distinct peak in the change of probability (ΔK) (Earl & vonHoldt, 2012). Discriminant analysis of principal components (DAPC) was complemented by the STRUCTURE analysis to further understand the population structure. DAPC is a multivariate analysis that functions with K-means and the selection method to infer and determine clusters using the level of genetic relatedness in the population (Sodedji et al., 2020). The optimum K was identified as the minimum number of clusters after which the Bayesian Information Criterion (BIC) decreases or increases by a negligible amount (Jombart et al., 2021; Sodedji et al., 2020). The DAPC, multivariate analysis was performed using the Adegenet package in R Software (Ref). The soybean accessions were assigned to subpopulations based on the membership probability high than 0.70 (Zavinon et al., 2020).

The Principal Coordinate Analysis (PCoA) is a distance-based model which uses jointly a dissimilarity matrix calculated with a simple-matching index. The PCoA of the DArT-seq markers was performed using PAST Software v.3.14 (Hammer et al., 2001). This software produces graphical representations on Euclidean plans which preserve at best the distances between units. The PCoA analysis was performed to complete the comprehension of the STRUCTURE and DAPC results.

An identity by state distance matrix was generated using TASSEL for phylogenetic relationship examination and clustering confirmation (Bradbury et al., 2007). The analysis of phylogenetic relationships among the genotypes was performed using a Euclidean distance matrix generated with Power Marker software v3.25 (K. Liu & Muse, 2005). The phylogenetic tree was constructed using the Neighbor-Joining algorithm and exported for visualization and annotation in MEGA-X software v10.18 (Newman et al., 2016) The coloration of genotype names was made according to the countries of origin.

Analysis of Molecular Variance (AMOVA) allows detecting population differentiation utilizing molecular markers (Excoffier et al., 1992). The AMOVA was performed with the Genetic Analysis in Excel (GenAIEx v.6.41) packages with the SNP markers and the repetition of the genotypes to the countries of their origin (Peakall & Smouse, 2006, 2012). Molecular data have been numerically coded (A=1, C=2, G=3, and T=4) as suggested in the GenAIEx manual (Blyton, MDJ; Flanagan, 2006). The coded dataset was run into GenAIEx for AMOVA with significance tested by 999 random permutations. The large genetic variation among the populations based on the countries of their origin was performed using phi-statistics. Both the variation among the population (PhiPR) and the variation within the population (PhiPT) was emitted.

Results

Summary statistics of the SNP information

The DArTseq generated SNP markers were 14,082 from 282 soybean accessions. A large number 7,687 (54.6%) were discarded after the filtering and imputation of the raw data. The remaining markers namely 6,395 SNPs representing 45.4% of the DArTseq generated SNP markers were used for the analysis. The 6,395 SNPs markers matched the criteria of the data to use for genetic diversity analysis. The 6395 SNPs were distributed across the 20 soybean chromosomes; chromosome number 12 and Chromosome number 18 are respectively with the lower and the higher concentration of SNPs (Figure 1). The diversity of the retained SNP markers is presented in table 1. The SNP markers diversity analysis revealed that soybean SNP had an average MAF (minor allele frequency) and an average polymorphic information content (PIC) respectively of 0.22 and 0.29. Gene diversity ranged from 0.16 to 0.51 with an average of 0.35. The means of expected and observed heterozygosity are respectively 0.33 and 0.05 (Table 1).

Population Structure

The comprehension of the population structure was investigated with the Bayesian Information Criterion (BIC) supplemented by the discriminant analysis of principal components (DAPC) and the principal coordinate analysis (PCoA). The information gathered from the determination of the populations at each K-value and membership coefficients (qi) in STRUCTURE analysis was very instructive (Figs 3A, 3B, and 3C). The simulation models of the logarithm probability relative to the standard deviation (ΔK) estimated from the 6,395 SNP markers presented a peak at ΔK = 2 (Fig 3A), which explained the optimum number of subpopulations. At ΔK = 2, Subpopulation I and Subpopulation II consisted of 105 soybean accessions (40.7%) and 173 soybean accessions (62.23%) respectively. The number of accessions from Uganda were high in the two subpopulations accounting for 86 (81.9%) in subpopulations 1 and 107 (61.9%) in subpopulations II. The accessions from Zimbabwe and Taiwan were respectively least represented in the subpopulation I (Figure 2, S2 Table). At ΔK = 3, the populations from Zimbabwe were dispatched in the subpopulations II (5 accessions) and III (3 accessions). We observed that populations from the USA that have been in subpopulation I remained in the same group at both ΔK = 2 and ΔK = 3, only the accessions in subpopulation II at ΔK = 2 dispatched in subpopulation II and III with respectively 13 and 6 accessions (Figure 2, S2 Table).

Discriminant analysis of principal components (DAPC) approach of population structure determination was further carried out to assess the subclusters. The curve of the Bayesian information criterion (BIC) in the DAPC method versus the number of clusters describes a quick decline from 1 to 3, followed by the BIC declining value reduction (Figure 4A). Therefore, K=3 is the suggested optimum number of clusters inferred by the DAPC. Based on the possibility of cluster membership assignment, the DAPC clustering (Figure 4B and Figure 4C) represented a good fit with the STRUCTURE at ΔK=3 with some deviation in the number of soybean accessions allocated in each subgroup (S2 Table). Group I consisted of 89 soybean accessions (32%) with the majority from Uganda (95.5%) and only 4 accessions (4.5%) from the USA. Groups II and III are varied regarding the population's origins, each has accessions from the 4 countries and regrouped respectively 84 (30.2%) and 105 (37.8%) accessions (Table 2).

A complementary analysis of the PCoA was used to understand the distance between soybean accessions on the Euclidian figure. The PCoA of the 278 soybean accessions based on 6395 SNPs markers showed a considerable genetic variability within and among the soybean populations. The distribution within the area defined by the PCo axes was not uniform. Therefore, the soybean accessions have a genetic structure (Figure 5A). The population is subdivided into two different sets with some outliers. The PCoA confirmed the result given by STRUCTURE at ΔK=2. The two subpopulations are easily distinguishable (Figure 5A). The subpopulation I haven't the soybean accessions from Zimbabwe, in green color (Figure 5A).

Hierarchical clustering of the Soybean germplasm

A phylogenetic tree was further generated. The 278 accessions were allocated to clusters by counting the number of branches at the first node in the dendrogram. At that point, each branch becomes a cluster considering the height of the others. The soybean populations split into three major clusters (clusters 1, 2, and 3) (Figure 5B). Two clusters 1 and 2 are the minority with respectively 39 and 10 soybean accessions and are exclusively from Uganda. All 219 remaining populations are in cluster 3 and develop other subclusters with some intermixture regarding the country of origin (Figure 5B).

Genetic divergences of subpopulations inferred by the Analysis of Molecular Variance

The Analysis of Molecular Variance (AMOVA) partitioned the genetic variance at two-level (among and within the subpopulations). Each level contributed to a varying degree of genetic variation to the total existing variation. This analysis was applied to two different considerations. First, based on countries of origin of soybean accessions, and secondly, based on the DAPC result. The genetic variance among and within the subpopulations was significant in both scenarios. Based on the countries of origin of the soybean accessions, the contribution of the variation among populations (5%) is less than the contribution of the variation within populations (95%) to the global variation of the populations (Table 3). The DAPC clustered the populations into 3 clusters. The AMOVA within and among the subpopulation from the DAPC reveals that the grouping is contributing 11% when the variation within populations is contributing 89 % of the total variation in the soybean population. Furthermore, the AMOVA based on the DAPC subpopulation showed a high level of variation (Phi-statistic = 0.114) than the analysis based on the countries of origins of the soybean accessions (Phi-statistic = 0.051) (Table 3).

Discussion

Molecular characterization of crop germplasm is essential for its efficient utilization (Adoukonou-Sagbadja et al., 2007; Carter et al., 2016). This study provides the first discernment of the genetic diversity and population structure in a large and representative collection of soybean in Uganda using SNP markers. The information generated from this study is beneficial not only for germplasm management and conservation of soybean but also for exploitation in the breeding.

In this study, molecular markers used were able to detect considerable genetic variability among the soybean accessions. The estimated diversity parameters in this study using 6,395 SNPs markers were higher than those estimated using DArT-Seq SNPs markers in others crops. The PIC of 0.21 and of 0.24 (lower than PIC = 0.29 obtained in this study) were reported respectively in perennial pasture *Phalaris aquatica* (Gapare et al., 2021) and in cowpea (Sodedji et al., 2020). In cowpea genetic analysis based on DArTseq, the expected heterozygocity of 0.12 (Ketema et al., 2020) and 0.30 (Sodedji et al., 2020) were reported. The expected heterozygocity of 0.33 reported in this study revealed that the diversity in the Uganda soybean accession is high and then, useable for soybean improvement.

In a KASP-markers-based analysis of the genetic diversity of soybean lines adapted to Sub-Saharan Africa, Chander et al. (2021) reported a mean genetic diversity and a PIC respectively of 0. 414 and 0.324. Those parameters are high than those revealed in this study. In a comparison study of the genetic diversity of the Chinese and the USA soybean accessions based on the Illumina SoySNP6K iSelect Bead chip genotyping tool, the PIC of 0.2643 and 0.2408 were respectively reported (Z. Liu et al., 2017); which is less than the PIC reported in this study. Some of the observed discrepancies could be attributed to the composition of the genotype set, the methods of revealing PCR products, and the genotyping model used in the study. In fact, there is no significant difference between the reported values of genetic parameters. All these results supported our conclusion that importantly large genetic diversity exists in the Ugandan soybean germplasm.

The structure of the Ugandan soybean germplasm was assessed. The results of this study revealed the existence of a genetic structure within the Ugandan soybean germplasm. Previously, the presence of a genetic population in African cultivated soybean was detected using SNPs markers (Chander et al., 2021; Tonny Obua et al., 2020; Shaibu et al., 2022). Indeed, the ΔK method described by Evanno et al. (2005) suggested the subdivision of the germplasm into two genetic subpopulations (Fig. 3B). This was also confirmed by the supplementary analysis of the PCoA (Fig. 5A). A similar structure was reported in soybean (Z. Liu et al., 2017; Tonny Obua et al., 2020; Shaibu et al., 2022). But, other values of ΔK are possible since some peaks were observed for ΔK (Fig. 3a). This hypothesis was supported by the DAPC (Fig. 4); which subdivided the soybean germplasm into three subpopulations. The clustering of the germplasm into 2 or 3 subpopulations could not qualitatively affect our conclusion, wherever other researchers reported the soybean population structure with more than 3 ($\Delta K \geq 3$) subpopulations (Chander et al., 2021; Jo et al., 2021)

The analysis of molecular variance revealed that the contribution to the genetic diversity of the variance within the populations is higher than that of the variane among the populations. Similar results have been reported in soybean (Shaibu et al., 2022; Zhao et al., 2018) and other crops such as rice (Aesomnuk et al., 2021), cowpea (Gomes et al., 2020; Ketema et al., 2020), potatoes (Lee et al., 2021), and in perennial pasture grass *Phalaris aquatic* (Gapare et al., 2021).

The Neighbor-Joining tree did not separate the soybean accessions of Uganda and those from the USA, Zimbabwe, and Taiwan. The soybean accessions from Uganda and other origins were clustered together and constitute the named cluster 3 (Fig. 5B). But, a group of soybean accessions from Uganda separated clearly from all other accessions and constituted two well-defined clusters. This high similarity in cluster 3 between the soybean accessions in the Ugandan germplasm indicates that these lines possibly shared a common genetic background or ancestries.

In a practical approach, this study on Ugandan soybean accession genetic diversity and its population structure is offering the predictions to improve the soybean characteristics. The different genetic parameters estimated, especially the high level of allelic diversity (Fig. 1, Table 1) indicate that the investigated soybean germplasm set up a rich resource that can be utilized by soybean breeders. In applied breeding, the comprehension of the population structure in the germplasm is crucial to identifying the genes and the quantitative traits loci that control the phenotypic variation (Zavinon et al., 2020). Therefore the presence of the genetic structure in the Uganda soybean germplasm which possesses a high variability in morphological traits gives it the qualities of good raw materials for soybean breeding.

Conclusion

This study provides a detailed insight into the genetic diversity and structure in the Ugandan soybean accessions using the 6395 high-quality SNPs markers. The mean value of genetic diversity and the polymorphism information content was respectively estimated at 0.35 and 0.29. The analysis of molecular variance reveals that the variance among the populations is lower than the variance within the populations. The soybean accessions could be classified into two groups based on the model-based population structure and the principal coordinate analysis or into three groups based on the discriminant analysis of the principal component and the Neighbor-joining tree. The information provided in this study will be helpful for breeders.

Declarations

Ethical Approval

Not applicable

Competing interests

The authors declare no competing interests

Authors' contributions: Conceptualization, M.M.L, T.G., and I.D.O.; Methodology, M.M.L, E.A.A. and I.D.O.; Formal analysis, M.M.L., and E.A.A.; Writing-original draft preparation, M.M.L; Writing-review, A.P., H.W., P.W., and M.O.S.; Editing paper: M.M.L, I.D.O., and E.A.A.; Supervision, P.T., and T.G. Funding acquisition, R.E. All authors have read and agreed to the published version of the manuscript.

Funding

The project was supported by the Makerere Regional Centre for Crops Improvement (MaRCCI) under the World Bank grant No 126974 (Africa Higher Education Centres of Excellence Project).

Availability of data and materials

All data that are not in this write up are available as supplementory material.

Acknowledgments: We thank Mr. Paul Kabayi from the National Crops Resources Research Institute (NaCCRI) of Uganda and Mr. Georges Iga from the Makerere University Soybean Research and Development program for assembling the soybean accessions. The Makerere Regional Centre for Crops Improvement (MaRCCI) and the World Bank for the funds.

References

1. Adoukonou-Sagbadja, H., Wagner, C., Dansi, A., Ahlemeyer, J., Daïnou, O., Akpagana, K., Ordon, F., & Friedt, W. (2007). Genetic diversity and population differentiation of traditional fonio millet (Digitaria

spp.) landraces from different agro-ecological zones of West Africa. Theoretical and Applied Genetics, *115*(7), 917–931. https://doi.org/10.1007/s00122-007-0618-x

2. Adu, G. B., Badu-Apraku, B., Akromah, R., Garcia-Oliveira, A. L., Awuku, F. J., & Gedil, M. (2019). Genetic diversity and population structure of early-maturing tropical maize inbred lines using SNP markers. PLoS ONE, *14*(4), 1–12. https://doi.org/10.1371/journal.pone.0214810
3. Aesomnuk, W., Ruengphayak, S., Ruanjaichon, V., Sreewongchai, T., Malumpong, C., Vanavichit, A., Toojinda, T., Wanchana, S., & Arikit, S. (2021). Estimation of the genetic diversity and population structure of thailand's rice landraces using snp markers. Agronomy, *11*(5), 1–14. https://doi.org/10.3390/agronomy11050995
4. Blyton, MDJ; Flanagan, N. (2006). *A comprehensive guide to GenAlEx 6.5*. Australian National University. https://biology-assets.anu.edu.au/GenAlEx/Download_files/GenAlEx 6.5 Guide.pdf
5. Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics, *23*(19), 2633–2635. https://doi.org/10.1093/bioinformatics/btm308
6. Carter, T. E., Nelson, R. L., Sneller, C. H., & Cui, Z. (2016). Genetic diversity in soybean. Soybeans: Improvement, Production, and Uses, *16*, 304–416. https://doi.org/10.2134/agronmonogr16.3ed.c8
7. Chander, S., Garcia-Oliveira, A. L., Gedil, M., Shah, T., Otusanya, G. O., Asiedu, R., & Chigeza, G. (2021). Genetic diversity and population structure of soybean lines adapted to sub-saharan africa using single nucleotide polymorphism (Snp) markers. Agronomy, *11*(3). https://doi.org/10.3390/agronomy11030604
8. Chander, S., Ortega-Beltran, A., Bandyopadhyay, R., Sheoran, P., Oluwayemisi Ige, G., Vasconcelos, M. W., & Garcia-Oliveira, A. L. (2019). Prospects for durable resistance against an old soybean enemy: A Four-Decade Journey from Rpp1 (Resistance to Phakopsora pachyrhizi) to Rpp7. Agronomy, *9*(7). https://doi.org/10.3390/agronomy9070348
9. Chiari, W. C., De Toledo, V. D. A. A., Ruvolo-Takasusuki, M. C. C., Braz De Oliveira, A. J., Sakaguti, E. S., Attencia, V. M., Costa, F. M., & Mitsui, M. H. (2005). Pollination of Soybean (Glycine max L. Merril) by Honeybees (Apis mellifera L.). Brazilian Archives of Biology and Technology, *48*(1), 31–36. https://doi.org/10.1590/S1516-89132005000100005
10. Clever, M., Phinehas, T., Mcebisi, M., Shorai, D., Isaac, O. D., Tonny, O., Hellen, K., & Patrick, R. (2020). Genetic diversity analysis among soybean genotypes using SSR markers in Uganda. African Journal of Biotechnology, *19*(7), 439–448. https://doi.org/10.5897/ajb2020.17152
11. Duran, C., Appleby, N., Edwards, D., & Batley, J. (2009). Molecular Genetic Markers: Discovery, Applications, Data Storage and Visualisation. Current Bioinformatics, *4*(1), 16–27. https://doi.org/10.2174/157489309787158198
12. Earl, D. A., & vonHoldt, B. M. (2012). STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources, *4*(2), 359–361. https://doi.org/10.1007/s12686-011-9548-7

13. 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. https://doi.org/10.1111/j.1365-294X.2005.02553.x

14. Excoffier, L., Smouse, P. E., & Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. Genetics, *131*(2), 479–491. https://doi.org/10.1093/genetics/131.2.479

15. Foster, R., Williamson, C. S., & Lunn, J. (2009). Culinary oils and their health effects. Nutrition Bulletin, *34*(1), 4–47. https://doi.org/10.1111/j.1467-3010.2008.01738.x

16. Fried, H. G., Narayanan, S., & Fallen, B. (2018). Characterization of a soybean (Glycine max L. Merr.) germplasm collection for root traits. PLoS ONE, *13*(7), 1–19. https://doi.org/10.1371/journal.pone.0200463

17. Fu, Y. B. (2015). Understanding crop genetic diversity under modern plant breeding. Theoretical and Applied Genetics, *128*(11), 2131–2142. https://doi.org/10.1007/s00122-015-2585-y

18. Gapare, W. J., Kilian, A., Stewart, A. V., Smith, K. F., & Culvenor, R. A. (2021). Genetic diversity among wild and cultivated germplasm of the perennial pasture grass Phalaris aquatica, using DArTseq SNP marker analysis. Crop and Pasture Science, *72*(10), 823–840. https://doi.org/10.1071/CP21112

19. Gomes, A. M. F., Draper, D., Talhinhas, P., Batista Santos, P., Simões, F., Nhantumbo, N., Massinga, R., Ramalho, J. C., Marques, I., & Ribeiro-Barros, A. I. (2020). Genetic diversity among cowpea (Vigna unguiculata (l.) walp.) landraces suggests central mozambique as an important hotspot of variation. Agronomy, *10*(12). https://doi.org/10.3390/agronomy10121893

20. Govindaraj, M., Vetriventhan, M., & Srinivasan, M. (2015). Importance of genetic diversity assessment in crop plants and its recent advances: An overview of its analytical perspectives. *Genetics Research International*, *2015*(Fig. 1). https://doi.org/10.1155/2015/431487

21. Gudeta, T. B. (2018). Molecular marker based genetic diversity in forest tree populations. Forestry Research and Engineering: International Journal, *2*(4), 176–182. https://doi.org/10.15406/freij.2018.02.00044

22. Gupta SK, M. J. (2017). Genetic diversity and population structure of Indian soybean (Glycine max (L.) Merr.) revealed by Simple Sequence Repeat markers. Crop Sciences and Biotechnology, *20*(3), 221–231. https://doi.org/10.1007/s12892-017-0023-0

23. Hammer, Ø., Harper, D. A. T., & Ryan, P. D. (2001). PAST: Paleontological Statistics Software Package for Education and Data Analysis PAST : PALEONTOLOGICAL STATISTICS SOFTWARE PACKAGE FOR EDUCATION AND DATA ANALYSIS Even a cursory glance at the recent paleontological literature should convince anyone tha. Palaeontologia Electronica, *4*(1), 1–9.

24. Hartman, G. L., West, E. D., & Herman, T. K. (2011). Crops that feed the World 2. Soybean-worldwide production, use, and constraints caused by pathogens and pests. Food Security, *3*(1), 5–17. https://doi.org/10.1007/s12571-010-0108-x

25. Jo, H., Lee, J. Y. J. D., Cho, H., Choi, H. J., Son, C. K., Bae, J. S., Bilyeu, K., Song, J. T., & Lee, J. Y. J. D. (2021). Genetic diversity of soybeans (Glycine max (L.) merr.) with black seed coats and green

cotyledons in Korean germplasm. Agronomy, *11*(3). https://doi.org/10.3390/agronomy11030581

26. Jombart, T., Solymos, P., Cori, A., Calboli, F., Kamvar, Z. N., & Lustrik, R. (2021). *Package ' adegenet '. Exploratory Analysis of Genetic and Genomic Data*. https://github.com/thibautjombart/adegenet
27. Ketema, S., Tesfaye, B., Keneni, G., Fenta, B. A., Assefa, E., Greliche, N., Machuka, E., & Yao, N. (2020). DArTSeq SNP-based markers revealed high genetic diversity and structured population in Ethiopian cowpea [Vigna unguiculata (L.) Walp] germplasms. PLoS ONE, *15*(10 October), 1–20. https://doi.org/10.1371/journal.pone.0239122
28. Khojely, D. M., Ibrahim, S. E., Sapey, E., & Han, T. (2018). History, current status, and prospects of soybean production and research in sub-Saharan Africa. Crop Journal, *6*(3), 226–235. https://doi.org/10.1016/j.cj.2018.03.006
29. 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. https://doi.org/10.1007/978-1-61779-870-2_5
30. Kujane, K., Sedibe, M. M., & Mofokeng, A. (2019). Genetic diversity analysis of soybean (Glycine max (L.) Merr.) genotypes making use of SSR markers. Australian Journal of Crop Science, *13*(7), 1113–1119. https://doi.org/10.21475/ajcs.19.13.07.p1638
31. Lee, K. J., Sebastin, R., Cho, G. T., Yoon, M., Lee, G. A., & Hyun, D. Y. (2021). Genetic diversity and population structure of potato germplasm in rda-genebank: Utilization for breeding and conservation. Plants, *10*(4). https://doi.org/10.3390/plants10040752
32. Liu, K., & Muse, S. V. (2005). PowerMaker: An integrated analysis environment for genetic maker analysis. Bioinformatics, *21*(9), 2128–2129. https://doi.org/10.1093/bioinformatics/bti282
33. Liu, Z., Li, H., Wen, Z., Fan, X., Li, Y., Guan, R., Guo, Y., Wang, S., Wang, D., & Qiu, L. (2017). Comparison of genetic diversity between Chinese and american soybean (Glycine max (L.)) accessions revealed by high-density SNPs. Frontiers in Plant Science, *8*(November). https://doi.org/10.3389/fpls.2017.02014
34. Macherey-nagel GmbH. (2018). *Genomic DNA from plant User manual NucleoMag ® Plant*.
35. Malik, M. F. A., Qureshi, A. S., Ashraf, M., Khan, M. R., & Javed, A. (2009). Evaluation of genetic diversity in soybean (Glycine max) lines using seed protein electrophoresis. Australian Journal of Crop Science, *3*(2), 107–112.
36. Mohammadi, S. A. and, & Prasanna, B. M. (2003). Analysis of Genetic Diversity in Crop Plants — Salient Statistical Tools. Crop Science, *43*, 1235–1248. https://doi.org/10.2135/cropsci2003.1235
37. Money, D., Gardner, K., Migicovsky, Z., Schwaninger, H., Zhong, G. Y., & Myles, S. (2015). LinkImpute: Fast and accurate genotype imputation for nonmodel organisms. *G3: Genes, Genomes, Genetics*, *5*(11), 2383–2390. https://doi.org/10.1534/g3.115.021667
38. Mukhopadhyay T. & Bhattacharjee S. (2016). Genetic Diversity: Its Importance and Measurements. In N. A. B. Aabid Hussain Mir (Ed.), *Conserving biological diversity: A multiscaled approach* (2016th ed., Issue October, pp. 251–295). Research India Publications.

39. Mutegi, J., & Zingore, S. (2013). Boosting soybean production for improved food security and incomes in Africa. African Journal of Agricultural Research, 1–8. http://ssa.ipni.net/ipniweb/region/africa.nsf/0/28600CA4712A18F 685257BE100695F27/$FILE/Soybean production in SSA BMPs, Challenges and Opportunities.pdf

40. Nadeem, M. A., Nawaz, M. A., Shahid, M. Q., Doğan, Y., Comertpay, G., Yıldız, M., Hatipoğlu, R., Ahmad, F., Alsaleh, A., Labhane, N., Özkan, H., Chung, G., & Baloch, F. S. (2018). DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing. Biotechnology and Biotechnological Equipment, *32*(2), 261–285. https://doi.org/10.1080/13102818.2017.1400401

41. Naik, S., Madhusudan, K., Motagi, B., Nadaf, H., & Thimmaraju, -. (2016). Diversity in soybean (Glycine max) accessions based on morphological characterization and seed longevity characteristics. Progressive Research – An International Journal, *11*(03), 377–381.

42. Newman, L., Duffus, A. L. J., & Lee, C. (2016). Using the free program MEGA to build phylogenetic trees from molecular data. American Biology Teacher, *78*(7), 608–612. https://doi.org/10.1525/abt.2016.78.7.608

43. Nkongolo, K., Alamri, S., & Michael, P. (2020). Assessment of Genetic Variation in Soybean (Glycine max) Accessions from International Gene Pools Using RAPD Markers: Comparison with the ISSR System. American Journal of Plant Sciences, *11*(09), 1414–1428. https://doi.org/10.4236/ajps.2020.119102

44. Obua, T., Nabasirye, M., Namara, M., Tusiime, G., Maphosa, M., & Tukamuhabwa, P. (2020). Yield stability of tropical soybean genotypes in selected agro-ecologies in Uganda. South African Journal of Plant and Soil, *37*(2), 168–173. https://doi.org/10.1080/02571862.2019.1678687

45. Obua, Tonny, Sserumaga, J. P., Opiyo, S. O., Tukamuhabwa, P., Odong, T. L., Mutuku, J., & Yao, N. (2020). Genetic Diversity and Population Structure Analysis of Tropical Soybean (Glycine max (L.) Merrill) using Single Nucleotide Polymorphic Markers. Global Journal of Science Frontier Research, *August*, 35–43. https://doi.org/10.34257/gjsfrdvol20is6pg35

46. Peakall, R., & Smouse, P. E. (2006). GENALEX 6: Genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecology Notes, *6*(1), 288–295. https://doi.org/10.1111/j.1471-8286.2005.01155.x

47. Peakall, R., & Smouse, P. E. (2012). GenALEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research-an update. Bioinformatics, *28*(19), 2537–2539. https://doi.org/10.1093/bioinformatics/bts460

48. Porras-Hurtado, L., Ruiz, Y., Santos, C., Phillips, C., Carracedo, Á., & Lareu, M. V. (2013). An overview of STRUCTURE: Applications, parameter settings, and supporting software. Frontiers in Genetics, *4*(MAY), 1–13. https://doi.org/10.3389/fgene.2013.00098

49. Raza, S., Shoaib, M. W., & Mubeen, H. (2016). Genetic Markers: Importance, uses and applications. International Journal of Scientific and Research Publications, *6*(3), 2250–3153. www.ijsrp.org

50. Roldán-Ruiz, I., Van Euwijk, F. A., Gilliland, T. J., Dubreuil, P., Dillmann, C., Lallemand, J., De Loose, M., & Baril, C. P. (2001). A comparative study of molecular and morphological methods of describing relationships between perennial ryegrass (Lolium perenne L.) varieties. Theoretical and Applied Genetics, *103*(8), 1138–1150. https://doi.org/10.1007/s001220100571
51. Shaibu, A. S., Ibrahim, H., Miko, Z. L., Mohammed, I. B., Mohammed, S. G., Yusuf, H. L., Kamara, A. Y., Omoigui, L. O., & Karikari, B. (2022). Assessment of the Genetic Structure and Diversity of Soybean and Single Nucleotide Polymorphism Markers. Plants, *11*(68). https://doi.org/https://doi.org/10.3390/plants11010068
52. Sinclair, T. R., Marrou, H., Soltani, A., Vadez, V., & Chandolu, K. C. (2014). Soybean production potential in Africa. Global Food Security, *3*(1), 31–40. https://doi.org/10.1016/j.gfs.2013.12.001
53. Sodedji, A. F. K., Agbahoungba, S., Agoyi, E. E., Kafoutchoni, K. M., Kim, H., Nguetta, S.-P. A., & Assogbadjo, A. E. (2020). DArT-seq based SNP analysis of diversity, population structure and linkage disequilibrium among 274 cowpea (Vigna unguiculata (L.) Walp.) accessions. *Research Square*, 1–19.
54. Tefera, H., Kamara, A. Y., Asafo-Adjei, B., & Dashiell, K. E. (2009). Improvement in grain and fodder yields of early-maturing promiscuous soybean varieties in the Guinea savanna of Nigeria. Crop Science, *49*(6), 2037–2042. https://doi.org/10.2135/cropsci2009.02.0081
55. Torres, A. R., Grunvald, A. K., Martins, T. B., Aparecida Dos Santos, M., Lemos, N. G., Silva, L. A. S., & Hungria, M. (2015). Genetic structure and diversity of a soybean germplasm considering biological nitrogen fixation and protein content. Scientia Agricola, *72*(1), 47–52. https://doi.org/10.1590/0103-9016-2014-0039
56. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics, *17*(6), 520–525. https://doi.org/10.1093/bioinformatics/17.6.520
57. Tukamuhabwa, P., Obaa, B., Obua, T., Namara, M., Okii, D., & Kabayi, P. (2016). *Status of Soybean Production and Impact Indicators of New Soybean Varieties in Uganda*. https://soybeanafrica.com/docs/Soybean Survey Report.Uganda.pdf
58. Zavinon, F., Adoukonou-Sagbadja, H., Keilwagen, J., Lehnert, H., Ordon, F., & Perovic, D. (2020). Genetic diversity and population structure in Beninese pigeon pea [Cajanus cajan (L.) Huth] landraces collection revealed by SSR and genome wide SNP markers. Genetic Resources and Crop Evolution, *67*(1), 191–208. https://doi.org/10.1007/s10722-019-00864-9
59. Zhao, H., Wang, Y., Xing, F., Liu, X., Yuan, C., Qi, G., Guo, J., & Dong, Y. (2018). The genetic diversity and geographic differentiation of the wild soybean in Northeast China based on nuclear microsatellite variation. *International Journal of Genomics*, *2018*. https://doi.org/10.1155/2018/8561458
60. Zimmer, Y. (2010). Competitiveness of rapeseed, soybeans and palm oil. Journal of Oilseed Brassica, *1*(12), 84–90.

Tables

Table 1: Profile of Single Nucleotide Polymorphism (SNP)

SNP Markers Profile	Mean	Min[a]	Max[b]
Major allele frequency	0.77	0.5	0.96
Minor allele frequency	0.22	0.04	0.50
Expected heterozygosity	0.33	0.09	0.50
Observed heterozygosity	0.05	0.01	0.58
Genetic diversity	0.35	0.16	0.51
Polymorphism Information Content	0.29	0.08	0.74

NB: [a]Minimum and [b]Maximum

Table 2: Membership clustering or grouping by discriminant analysis of principal components

Subpopulations	**Genotypes**	**%**	**Origin**			
			Uganda	USA	Taiwan	Zimbabwe
I	89	32	85	4	0	0
II	84	30.2	41	22	17	4
III	105	37.8	66	13	21	5
TOTAL	278	100	192	39	38	9

Table 3: Analysis of molecular variance showing the partitioning of genetic variation within and among 276 soybean accessions

Source	df	SS	MS	Est. Var.	%	Phi-statistic	P-value
AMOVA among/within 4 groups of populations based on the soybean origin							
Among Pops	3	10984.1	3661.3	57.8	5%		
Within Pops	272	292526.4	1075.5	1075.5	95%	0.05	0.001
Total	275	303510.5		1133.3	100%		
AMOVA among/within 3 groups of soybean populations based on DAPC analysis							
Among Pops	2	32246.8	16123.400	162.1	11%		
Within Pops	273	345501.2	1265.572	1265.5	89%	0.11	0.001
Total	275	377748.0		1427.7	100%		

Figures

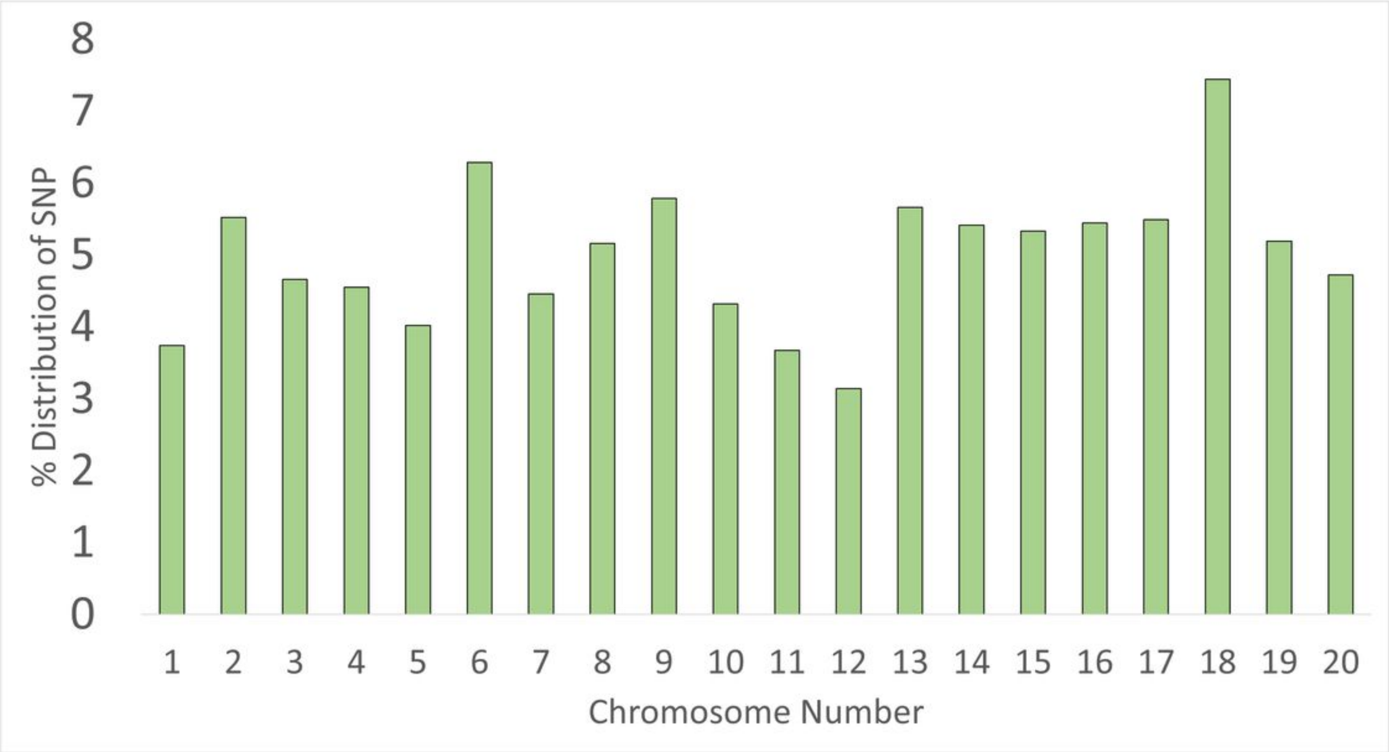

Figure 1

Distribution (in percentage) of Single Nucleotide Polymorphism (SNP) Markers on the twenty chromosomes of soybean

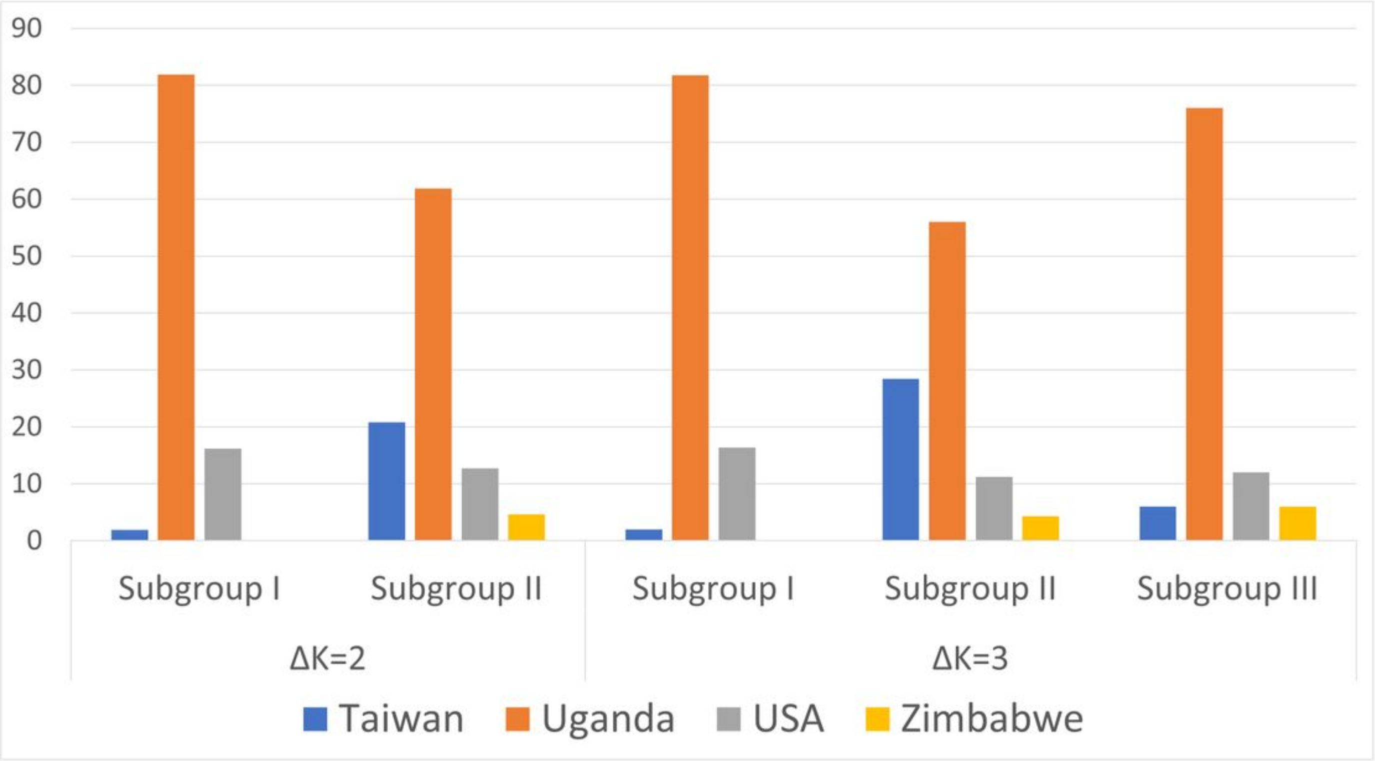

Figure 2

Percentage of soybean accessions assigned to subgroups considering the country of origin at ΔK= 2 and ΔK=3

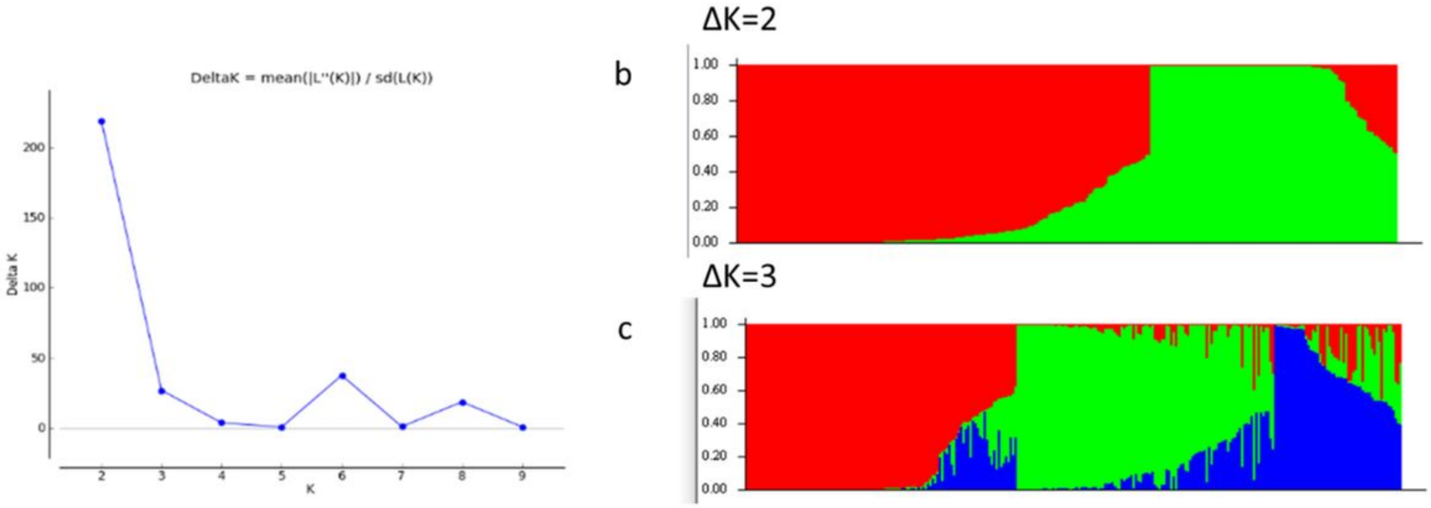

Figure 3

Likelihood of ΔK showing the K value (a). The structure graphs of K=2 (b) and K=3 (c) of 278 soybean accessions assigned respectively into two and three clusters

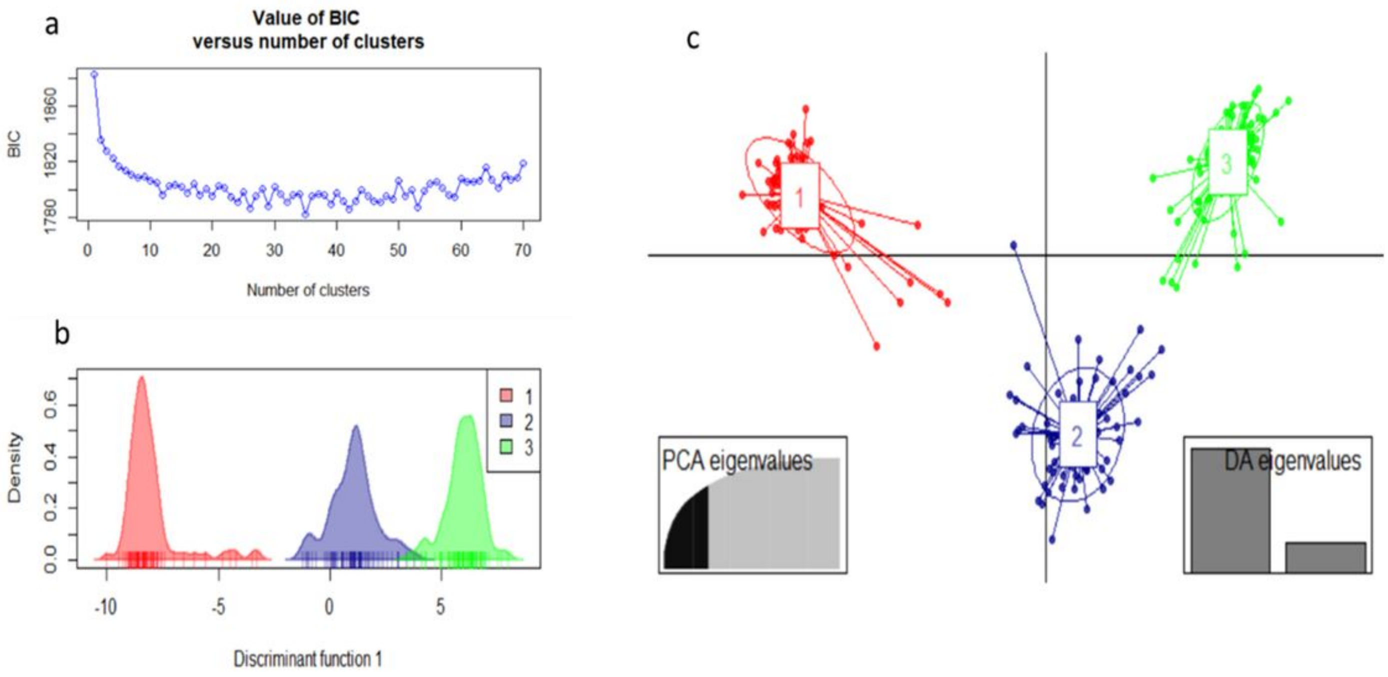

Figure 4

Bayesian information criterion (BIC) showing a rapid decline from 1 to 3 followed by BIC declining value reduction at 3 (a), Soybean accessions composition plots using densities of individuals, partitioned into 3 groups by DAPC (Discriminant Analysis 1). Group I is further apart from Group II and III. Groups II and III present an intersected area (b) and DAPC scatters plot of 278 soybean accessions shows three separated groups (c)

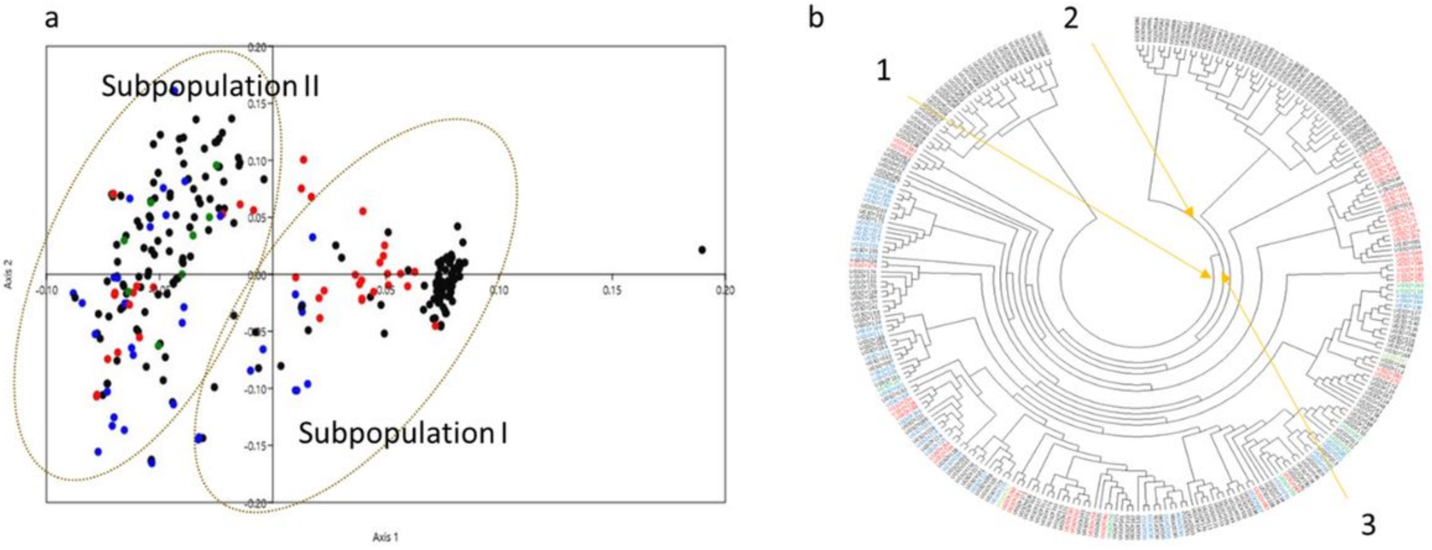

Figure 5

Principal coordinate analysis (a) and Neighbor-Joining tree (b) of the 278 soybean accessions. Soybean accessions from Uganda, USA, Taiwan and Zimbabwe are shown in black, red, blue and green respectively

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- S1Table.docx
- S2Table.docx