STUDY ON GRAIN YIELD STABILITY, MOLECULAR DIVERSITY AND MULTI-TRAIT RELATIONSHIPS AMONG ELITE SOYBEAN LINES

By

Bertha Mwayi Kachala

BSc in Horticulture (Malawi)

A dissertation submitted in partial fulfilment of the academic requirements for Master of Science degree in Plant Breeding

School of Agricultural, Earth and Environmental Sciences

College of Agriculture, Engineering and Science

University of KwaZulu-Natal

Pietermaritzburg

Republic of South Africa

January 2018

GENERAL ABSTRACT

The demand for soybean production has increased in recent years, due to its multipurpose use for human food, livestock feed and industrial purposes. The soybean crop is one of the important source of oil and protein of the world, and is used as a source of high quality edible oil and protein. For a quantitative trait, yield is known to be influenced by changes in the environment in which the crop is grown, suggesting the need to evaluate soybean lines in different growing regions to assess their adaptability and stability. In plant breeding, selection is one of the most important stages in the breeding cycle. Multi-location testing of soybean genotypes precedes selection while genetic characterisation of germplasm enhances selection due to the variation realised and it is the basis for genetic improvement. The objectives of the study were: 1) to determine yield stability and adaptability of elite soybean lines across six locations, 2) to study genotype by trait associations and multiple trait relationships among soybean elite lines across six locations and 3) to assess the level of genetic diversity among the soybean elite lines using single nucleotide polymorphisms (SNP) markers.

The stability and adaptation study was carried out to investigate genotype by environment interaction (GEI) for grain yield of 26 elite soybean lines along with four checks grown in 6 environments spreading across three countries (Zambia, Malawi and Mozambique) in a 6 x 5 alpha lattice design. The additive main effect and multiplicative interaction model (AMMI) indicated that environments, genotypes and GEI significantly affected grain yield ($P<0.001$) and contributed 3.8%, 17% and 78%, respectively, to the total variation. Three AMMI interaction principal components (IPCA1, IPCA2 and IPCA3) were significant ($P<0.01$). Genotype plus GEI (GGE) biplots were created based on the first two principal components, PC1 and PC2, which accounted for 39.23% and 26.86% of genotype plus GEI variation, respectively. The GGE biplot analysis ranked the genotypes for yield and stability, and environments for representativeness and discriminativeness. The relationships between genotypes and environments were also demonstrated. Genotype TGX 2001-3FM was identified as the ideal genotype with high yield mean performance and high stability. Therefore, it could be recommended for cultivar release if the study can be repeated to verify these findings. Chitedze in Malawi was the most informative test environment hence it is ideal for selecting generally adapted genotypes. Genotypes TGX 2002-4FM and TGX 2001-15DM were low yielding but with high stability hence can be recommended for further improvements.

For the second objective, a study was conducted using 30 genotypes to determine the correlation and path coefficient of secondary traits on yield. The genotype by trait biplot is a tool that graphically compares genotypes on the basis of multiple traits and graphically visualises trait relationships, and genotype-trait associations. Trait profiling of genotypes through genotype-trait association analysis helps in making decisions on which genotypes to use as parents for a breeding programme. Significant differences among genotypes were observed for all studied traits. Correlation coefficient analysis presented that grain yield had a significant and negative correlation with days to 50% flowering. However, grain yield had a significant and positive correlation with plant height. Path coefficient analysis indicated that plant height and early vigour had a positive direct effect on yield while days to 50% flowering and days to 50% podding had negative indirect effects on yield via days to maturity. The genotype by trait biplot graphically showed consistent trait relationships and identified TGX 2001-3FM, TGX 2001-26DM and TGX 2002-3DM as genotypes that can be used as parents in breeding programmes for yield improvement.

Estimation of genetic diversity among 48 soybean lines from the International Institute for Tropical Agriculture (IITA) was conducted using 348 SNP markers. The average gene diversity and genetic distance ranged from 0.42 to 0.55 with an average of 0.47 and 0.61 to 0.87, respectively. The polymorphic information content ranged from 0.44 to 0.50 with a mean of 0.48. Genotypes TGX 2002-3DM and TGX 2002-3FM had the highest genetic distance between them indicating that they were highly diverse. The AMOVA indicated highly significant differences at F=0.001 with among individuals, among populations and within individuals contributing 45%, 28% and 26%, respectively. The 48 soybean lines were clustered in three main groups. The study indicated that genetic diversity exists among the IITA tested lines. The information obtained from the study, can be fully utilised in future soybean breeding programmes through crossing of diverse parents in order to incorporate new alleles to develop improved cultivars.

In general, the study showed the existence of genotype by environment of soybean grain yield across the selected locations in southern Africa. Based on the SNP markers, the study confirmed the existence of wide genetic diversity among the soybean lines. The lines with superior performances can be used for future breeding programmes or recommended for cultivar release.

DECLARATION

I, Bertha Mwayi Kachala, declare that

1. The research reported in this dissertation, except where otherwise indicated, is my original research.

2. This dissertation has not been submitted for any degree or examination at any other university.

3. This dissertation does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.

4. This dissertation does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:

   a. Their words have been re-written but the general information attributed to them has been referenced.

   b. Where their exact words have been used, then their writing has been placed in italics and inside quotation marks, and referenced.

5. This dissertation does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the dissertation and in the references sections.

Signed

Kachala

Bertha Mwayi Kachala

As the candidate's supervisors, we agree to submission of this dissertation:

Sibiya

Dr Julia Sibiya (Supervisor)

Dr Godfree Chigeza (Co-Supervisor)

ACKNOWLEDGEMENTS

First of all, I would like to give my praises and thanks to God, for giving me the wisdom and strength to carry on and for making my dream of finishing my Master's degree come true. This study was made possible through a scholarship from the Alliance for a Green Revolution in Africa (AGRA), many thanks for the opportunity.

My special gratitude goes to my supervisor and manager for the MSc Plant breeding programme, Dr Julia Sibiya, my role model, for the guidance, encouragement and advice during the course of study. I would like to extend my appreciation to Dr Cousin Musvosvi and Dr Amelework Beyene Assefa for their contribution towards the completion of this study. They played a vital role in developing and designing research concepts, provided moral and academic support, contributed significantly in conceptualization and making the ideas real.

My sincere appreciation goes to my co-supervisor and mentor Dr Godfree Chigeza of the International Institute of Tropical Agriculture (IITA) for his precious time, theoretical and practical inputs, support and constant guidance throughout the period of study. I am also thankful to him for providing the materials used in this study and deeply grateful to Miss C Nanchilima and the entire IITA SARAH team for their help and support during my entire research and internship work.

I would also like to extend many thanks to our administrative officer Ms. Andile Mshengu and the former, Ms. Jayshree Singh for the administrative support and the logistics to ensure the success of the study. I would like to express many thanks to my fellow classmates, MSc 2016 cohort for their love and encouragement all the way.

Lastly, I thank my family for their love, continuous prayers and unwavering support and encouragement throughout the entire study period. My siblings, Victoria, Pilirani and Yamikani for their emotional support, I am humbled.

DEDICATION

I would like to dedicate this thesis to my loving parents Stanzio and Margaret Kachala and my niece Mukuzike, my greatest cheerleaders, with lots of love.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AFLP = Amplified fragment length polymorphism

AMMI = Additive main effect multiplicative interaction

ANOVA = Analysis of variance

CV = Coefficient of Variation

DF = Degrees of freedom

DFFL = Days to fifty percent flowering

DM = Days to maturity

DPD = Days to fifty percent podding

EV = Early vigour

FAO = Food and agriculture organisation

GGE = Genotype main effects and genotype by environment interaction

GEI = Genotype by environment interaction

GT = Genotype by trait

GY = Grain yield

H_o = Observed genetic diversity within genotypes

H_e = Average gene diversity within genotypes

HSW = Hundred seed weight

IITA = International Institute for tropical agriculture

IPCA = Interaction principal component axis

KASP = Kompetittive allele specific polymerase chain reaction

Kg/ha = Kilograms per hectare

LSD = Least significance difference

MS = Mean squares

N_a = Total number of alleles per locus

N_e = Number of effective alleles per locus

PCR = Polymerase chain reaction

PIC = Polymorphic information content

PLHT = Plant height

RAPD = Random amplification of polymorphic DNA

RFLP = Restriction fragment length polymorphism

SS = Sum of squares

SNP = Single nucleotide polymorphisms

SSA = Sub – Saharan Africa

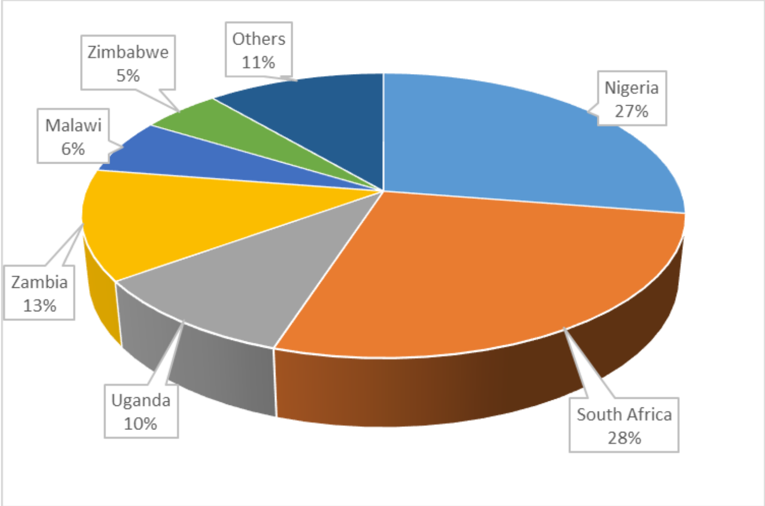
SSR = Microsatellites/ simple sequence repeats

UK = United Kingdom

INTRODUCTION

1 Background/ Justification

Soybean (*Glycine max* L.), is one of the most important protein and oil seed crops throughout the world. It is a leguminous crop that belongs to the *Fabacae* family and subfamily *Papilionoidae.* Soybean is believed to have been first domesticated in China around the 11th century and later spread to America and Africa. The world's leading soybean producers are the United States of America, which produces 32% of the total world's soybean followed by Brazil contributing 28%. Argentina, China and India contribute 21%, 7% and 4%, respectively, to the total global production. Africa contributes 1.3% of the total world production, with South Africa as the leading soybean producer followed by Nigeria and then Uganda (Figure 1). South Africa is the largest producer in the southern part of Africa dominating in both demand and production. Zambia is the second largest producer and exporter to countries such as Zimbabwe and Botswana (Abate *et al.,* 2012). Over the years, from 2010 to 2014 (Table 1), there has been an increased production in Zambia and Malawi implying that more farmers are investing in soybean production. With the involvement of International NGOs and the government, Malawi and Mozambique have rapidly increased their soybean production.



Source: Data from FAOSTAT, (2016) and calculations by author

Figure 1.1 Soybean production in Africa

Table 1 Soybean production from 2010 to 2014 in Malawi and Zambia

Country	Year	Harvested area (ha)	Quantity (tonnes)	Yield (t/ha)
Malawi	2010	75186	73356	0.98
	2011	75839	75665	1.00
	2012	102167	106592	1.04
	2013	114369	111977	0.98
	2014	139005	120903	0.87
Zambia	2010	60777	111887	1.84
	2011	59988	116539	1.94
	2012	84809	203038	2.39
	2013	124858	261063	2.09
	2014	113759	214179	1.88

Source: (FAOSTAT, 2016)

Soybean has multiple uses and benefits. It is the world's second leading source of oil and proteins (Gurmu *et al.,* 2009). The seed contains about 40% protein, 20% oil and a high calorie value (Singh and Shivakumar, 2010) making it an essential source of food and livestock feed. The soybean meal is rich in phosphorus, iron and calcium, making it perfect for animal feed. Additionally, it can be used as a raw material for industrial uses. It is the most common legume crop with other agronomic importance besides grain production. It helps in improving the soil fertility by capturing atmospheric nitrogen and fix it in the soil through symbiosis with rhizobia bacteria (Kumudini, 2010). When intercropped with other crops such as cereals and cassava, it has the potential to disturb life cycles of several pests, diseases and weeds like *Striga hermonthica*.

Soybean production is affected by various biotic and abiotic factors that lead to low yields of <1 t/ha in Africa (FAOSTAT, 2016). Among the biotic factors, diseases such as rust, bacterial blight, and pests including armyworm and beetles significantly reduce yield. Abiotic factors include drought and poor soil fertility. Varieties that are adapted to abiotic and biotic stresses in the soybean growing areas of the southern part of Africa are not known. Hence, evaluation of genetic diversity in soybean lines is essential for improvement of both yield and quality.

Plants are grown in areas that have different climatic and environmental attributes that includes temperature, rainfall, soil type, soil nutrients and cultural practices. Most soybean cultivars that are being produced have been genetically improved intensely for high grain yield (Das, 2005). For these cultivars to fully express their genetic potential for grain yield they require specific environmental conditions, hence the performance of these cultivars are different depending on the area they are being produced. This relative change in performance of cultivars across various environments is termed genotype by environment interaction (GEI).

The concept of global climatic change that is going on over the years is somehow responsible for altering the crop production environment (Acquaah, 2007). The problems that climate change will influence agriculture can be mitigated through the intervention by agricultural scientists developing ways to alleviate the impacts. Hence, plant breeding programmes need to engage in strategies that can help to adopt environment specific approaches to crop improvement (Reynolds *et al.,* 2001).

Stability is a concept that in most cases is a challenge in breeding programmes. Breeders are interested in good performing cultivars over a range of environments. However, there are GEI effects making cultivars have a high mean yield in other environments and a low mean yield in other environments, or showing better mean performance across environments. However, few genotypes may have average yield that is stable over wider environments (Cooper *et al.,* 2006). Knowledge of the pattern and magnitude of GEI and stability analysis is important for understanding the response of different genotypes to varying environments. Secondly, knowing the magnitude and patterns of GEI can be used in identifying superior soybean genotypes under the target environment and agronomic conditions. This will help to maximize specific adaptation and reducing the time to transfer new cultivars to growers from breeders (Cooper and Hammer, 1996). In plant breeding programmes, the common goal is also to identify traits that positively contribute to high yield. Therefore, it is critical to study traits in a crop and identify those that contribute to the trait of interest (Kinfe *et al.,* 2015). These trait profiles and associations identify their strengths and weaknesses and can be used in selection of parents in a breeding programme (Yan and Frégeau-Reid, 2008).

Soybean genotypes have been released from breeding programmes under different agro-climatic conditions by selection, hybridisation, introduction and mutation of soybean elite lines through systematic breeding programmes and evaluations. Genetic diversity among the cultivars is crucial to breeding programmes in germplasm enhancement and cultivar development (Dong *et al.,*, 2004). Several methods have been used to assess diversity in crops; this includes morphological, pedigree and biochemical markers. Molecular markers; simple sequence repeats (SSR), amplified fragment length polymorphisms (AFLP), random amplification of polymorphic DNA (RAPD) have also been used in quantifying the diversity in soybean but there is little information on studies which used single nucleotide polymorphisms (SNPs). Therefore, the knowledge gained from a genetic diversity in soybean elite lines can be used in future breeding programmes through selection of diverse genotypes as parents for soybean improvement.

In order to fully utilise the soybean elite lines bred by the International Institute of Tropical Agriculture (IITA), evaluations on yield stability and characterisation of lines using molecular markers can help to realise the potential in the lines that can used as a starting point for germplasm improvement and enhancement. Hence, the study was designed to assess the yield stability, multi-trait relationships and define the level of genetic diversity in the elite soybean lines from IITA.

2 Objectives

The overall goal of the study is to generate information that is crucial for soybean breeding programmes through identification of high yielding and stable cultivars and defining the level of diversity among the tested lines. The specific objectives were:

i. To determine yield stability and adaptability of elite soybean lines across six locations
ii. To understand the genotype by trait associations and multiple trait relationships among the soybean elite lines across six locations
iii. To assess the level of genetic diversity among the soybean elite lines using SNP markers

3 Research hypothesis

i. There is considerable genetic variability among soybean accessions based on molecular markers
ii. The performance, yield stability of the soybean elite lines are affected by genotype x environment interaction
iii. Genotype by trait associations and multiple trait relationships affect the performance of soybean elite lines

4 Dissertation outline

The dissertation is organised into five chapters following a journal paper format. As a result, there is some unavoidable repetition in the references and some overlaps in the introductory information between chapters. The referencing format is based on the Crop Science journal style. The outline of the dissertation is as shown below:

Introduction to thesis

Chapter 1: Literature Review

Chapter 2: Yield stability and adaptation analysis of elite soybean lines across diverse environments in Southern Africa

Chapter 3: Assessment of genetic diversity in tropical soybean lines using single nucleotide polymorphisms

Chapter 4: Analysis of soybean elite lines using genotype by trait, correlation and path coefficient

Chapter 5: General overview of the study

REFERENCES

Abate, T., A.D. Alene, D. Bergvinson, B. Shiferaw, S. Silim, A. Orr, et al. 2012. Tropical grain legumes in Africa and south Asia: knowledge and opportunities. International Crops Research Institute for the Semi-Arid Tropics.

Acquaah, G. 2007. Principles of plant breeding and genetics. Malden, MA USA: Blackwell Publishing.

Cooper, M. and G.L. Hammer. 1996. Plant adaptation and crop improvement. IRRI.

Cooper, M., F. van Eeuwijk, S.C. Chapman, D.W. Podlich and C. Löffler. 2006. Genotype-by-environment interactions under water limited conditions. In Drought adaptation in cereals (pp 51-96). Haworth press

Dong, Y., L. Zhao, B. Liu, Z. Wang, Z. Jin and H. Sun. 2004. The genetic diversity of cultivated soybean grown in China. Theoretical and Applied Genetics 108: 931-936.

FAOSTAT, 2016. Agriculture Organization of the United Nations Statistics Division (2014). Production Available in: http://faostat3. FAO. org/browse/Q/QC/S [Review date: April 2015].

Gurmu, F., H. Mohammed and G. Alemaw. 2009. Genotype x environment interactions and stability of soybean for grain yield and nutrition quality. African Crop Science Journal 17(2): 87-99

Kinfe, H., G. Alemayehu, L. Wolde and Y. Tsehaye. 2015. Correlation and Path Coefficient Analysis of Grain Yield and Yield Related Traits in Maize (*Zea mays* L.) Hybrids, at Bako, Ethiopia. Journal of Biology, Agriculture and Healthcare 5: 15-24.

Kumudini, S. 2010. Soybean growth and development. The Soybean: Botany, Production and Uses. British Library, London, UK: pp48-73.

Reynolds, M., S. Nagarajan, M. Razzaque and O. Ageeb. 2001. Heat tolerance. Application of physiology in wheat breeding: African Crop Science Journalpp124-135.

Singh, G. and B. Shivakumar. 2010. The role of soybean in agriculture. The Soybean: Botany, Production and Uses. CAB International, Oxfordshire, UK: pp24-47.

Vijendra Das, L. 2005. Genetics and Plant Breeding Revised. New Age International (P) Limited: India: pp34-48

Yan, W. and J. Frégeau-Reid. 2008. Breeding line selection based on multiple traits. Crop Science. 48: 417-423.

CHAPTER 1

LITERATURE REVIEW

1.1 Origin and distribution

Soybean (*Glycine max L*) is a self-pollinating leguminous crop, that has <1% outcrossing. It is diploid with 20 chromosome pairs (2n=40). The genus *Glycine* has two subgenera, *Soja* and *Glycine*, the sub genus *Soja* has two species namely *Glycine max* and *Soja seib.* Soybean is believed to have been derived from two wild progenitors *G. ussuriensis* and *usd* that are commonly found in East Asia. The crop was domesticated around 11th century in China, and later spread to neighbouring countries such as Mongolia and Japan 3000 years ago (Singh, 1991). China is known to be the centre of origin and diversity.

In most countries, soybean is grown as a commercial crop. The United States is the largest producer in the world, producing 38% of the soybean globally. It is followed by Brazil (26%), Argentina (21%), China (7%), India (4%) and Africa covers 1% of the global production (FAOSTAT, 2016). In Africa, the Chinese missionaries introduced soybean in the 19th century. Currently, the leading producer in Africa is South Africa (617 000 tons) which is followed by Nigeria (430 000 tons) (Abate *et al.*, 2012).

1.2 Botany

Soybean is an erect, annual plant that has dense green leaves covered with fine hairs. The first leaves are simple and grow opposite each other on the stem while the leaves that form subsequently are trifoliate. They have small flowers that consist of five separate; unequal petals that can vary in colour but are commonly violet or white. However, the morphology is diverse depending on the cultivar (Johnson and Bernard, 1962). The flowers and lateral branches form at the auxiliary buds at the point of contact between the leaf petiole and the main stem. The height of the plants can range from about 0.3 to 3.0 m.

The seed is made up of two parts, the seed coat, which covers and protects the embryo and two cotyledons that form part of the embryo region. The bean is attached to the pod at the hilum (Kumudini, 2010). Soybean seeds occur in various sizes, and in many seed coat colours, including black, brown, blue, yellow, green and mottled. Varieties differ in hilum colour and can be yellow, imperfect yellow, grey, buff, brown, black or imperfect black. Yellow hilum/ clear hilum soybeans with large seed size and thin but strong seed coat that is free from cracking and discoloration are preferred (Gandhi, 2009). However, the yellow and green seeds are more common.

Soybean pods are straight and sometimes slightly curved, reaching 20-70 mm long depending on the cultivar and environment, and they form in clusters of 1 - 9. Young pods are green in colour, covered in fine transparent hairs and when they mature, they are also hairy and range in colours such as brown or tan, black and yellow. This colour change happens as the plant's leaves turn yellow and fall off. The pods may contain 1-4 seeds depending on the cultivar (Krisnawati and Adie, 2015).

The root system of soybean consists of a taproot, which can grow up to 1.2 m into the soil. Furthermore, there is a proliferation of secondary roots that are arranged in four rows along the taproot. Most of the effective roots are found in the top 600 mm of soil; therefore, the soybean plant is a shallow feeder (Kumudini, 2010).

1.3 Importance of soybean

Soybean is an economically important leguminous crop that is grown for its oil and protein products (Tefera *et al.*, 2009). The soybean seed contains an average of 40% protein and 20% oil that is used for producing food products such as soymilk, soy flour, soy sauce and tofu (Fabiyi, 2006). It is also an important source of proteins in feed supplements for livestock. Besides its nutritive value, soybean has medicinal properties due to high iso-flavones content that reduce blood cancer, osteoporosis, blood cholesterol and heart diseases in human beings (Pathan and Sleper, 2008).

The crop also helps to improve soil fertility through biological nitrogen fixation, thus reducing the cost of purchasing inorganic fertilizers by resource constrained farmers (Misiko *et al.*, 2008). Soybean dual-purpose varieties have revealed its potential in reducing the levels of *Striga hermonthica* infestations when they are in rotational system with cereals (Chianu *et al.*, 2006). Furthermore, soybean is used as a raw material in industries for production of biodiesel, cosmetics, pesticides, hydraulic fluids, lubricants, paint removers and plastics; hence, smallholder farmers can utilize it as a beneficial crop for income generation (Pathan and Sleper, 2008).

Soybeans contain three lipoxynase isozymes that play a role in the development of beany off-flavour in food containing soy-protein that is unpleasant to some consumers. The off-flavour is caused by oxidation of polyunsaturated fatty acids (Wilson, 1996). The poor stability and off-flavours of soybean oil and protein products can be reduced by eliminating lipoxygenases from soybean seed (Reinprecht *et al.*, 2011). Some varieties that are lipoxygenase free have been developed and are referred to as “triple null” soybeans. These are highly preferred and normally used for edible soy-products such as soymilk and tofu because of less saturated fat resulting in healthier oil that is used for salad dressing and other food products.

1.4 Selection in soybean and important traits

Soybean is primarily bred for improved yield, high oil and protein content, pests and disease resistance, lodging resistance, drought tolerance, resistance to pod shattering and degree of biological nitrogen fixation (Tefera, 2011). Over the past 30 years, the role of soybean in industry as a source of protein and oil has developed significantly leading to the production of complex products (Cianzio *et al.*, 2007). Therefore, yield as well as protein and oil content are important to food and oil industries. Soybean varieties are distinguishable by various characteristics such as flower colour, pubescence colour, pod colour, seed colour, leaf shape and stem type among others. In the breeding process, these traits are selected at the target environments. Maturity is another important trait when selecting a soybean cultivar; however, most farmers prefer early maturing varieties (Duxburg *et al*., 1990).

Soybean is self-pollinating with a 1% chance of outcrossing. Due to this state, breeding methods such as pedigree breeding, single seed descent, bulk breeding and backcrosses have been used to develop new varieties (Miladinović *et al.*, 2015). These methods involve making crosses by hand pollination to produce hybrids, selection follows and ultimately the release of a superior cultivar (Burton and Miranda, 2013).

Of all soybean traits, 100-seed weight, protein content, and oil content are the most valuable phenotypic characteristics related to soybean seed quality (Borrás *et al.*, 2004). These soybean traits have variations that are genetically controlled; however, they are largely influenced by climatic and environmental conditions (Burton *et al.*, 2006). Soybean traits could vary greatly with geographical locations. For example, in the north western area of the United States, soybeans have higher seed oil content and lower seed protein content than those found in the south eastern states (Breene *et al.*, 1988)

Seed weight, plant height, protein content, and oil content could vary largely with variations in temperature. For example, in a controlled environment, 100-seed weight increased with increasing temperature to an optimum level (Sionit *et al.*, 1987) and then dropped (Baker *et al.*, 1989). Gibson and Mullen (1996) reported a similar temperature impact on seed oil content that was positively correlated with temperature until the higher end of the optimum range, whereas, the protein content indicated a negative response (Dornbos and Mullen, 1992).

1.5 Production constraints

A number of biotic and abiotic factors affects soybean production. The average grain yield in Africa is still low (<1 t/ha) (FAOSTAT, 2016) mostly because many growers have not accessed the improved varieties and some have no interest in soybean production because they are not aware of how to prepare it for consumption and postharvest handling techniques (IITA, 2009.)

Soybean production is also constrained by a number of biotic stresses, which include; soybean rust (*Phakopsora pachyrhizi* L), bacterial pustule (*Xanthomonas campestris* pv. *glycines* L), bacterial blight (*Pseudomonas amygdali* pv. *glycinea* L), frogeye leaf spot (*Cercospora sojina* L*),* red leaf blotch (*Phoma glycinicola* L) and soybean mosaic virus disease. The major insect pests affecting soybeans are armyworm (*Pseudaletia unipuncta* L), saltmarsh caterpillar (*Estigmene acrea* L), soybean looper (*Pseudoplusia includes* L), bean leaf beetle (*Cerotoma trifurcate* L), blister beetles (*Epicauta funebris*, *Epicauta vittata* L) and velvet bean caterpillar (*Anticarsia gemma* L) (Catchot, 2010).

Abiotic stresses include drought, flooding, salinity and nutrient deficient soils. Kehlenbeck *et al.* (1994) reported that annually, crop losses reach 42% due to abiotic stresses and crops resistant to these stresses are needed.

1.6 Genotype by environment interaction and stability

Several crops have been widely exposed to the genotype by environment interactions (GEI) studies (Alberts, 2004; Cooper *et al.,* 2006; Gurmu *et al.,* 2009). Genotype by environment interaction occurs when a few or more genotypes are tested across various environments and they have different responses to the environmental conditions. Thus, GEI is the differential response of genotypes to changes in the environment (Matter and Caligari, 1976). The consequence of the phenotypic variation depends largely on the environment. The variation is further complicated by the fact that not all the genotypes react in a similar way to changes in environment and no two environments are the same. Genotype by environment interaction is an important concept in plant breeding programmes because it delays progress from selection in any given environment (Yau, 1995). The phenotype of an individual is determined by the effect of the genotype and the environment surrounding it. Therefore an understanding of the genotypic and environmental causes of GEI is important at all stages of plant breeding, including the design of ideotypes, selection of parents based on traits and selection based on yield (Yan and Hunt, 1998).

The understanding of GEI in plant breeding programmes is important for improving the genotypes for higher yields (Alberts, 2004). The occurrence of GEI in multi-location evaluation trials leads to the selection of genotypes that perform among the best in one environment while they perform poorly in another test environment (Bekheit, 2000). This is the main hindering factor in selecting genotypes with a wide adaptability. GEI could be detected through a simple joint analysis of variance among trials repeated in more than one location and correct information about each location variation and genotype performance can be obtained (Crossa, 1990).

Soybean production increase is guaranteed when high yielding and early maturing cultivars across a wide range of different environments are developed. A knowledge of the genetic variability present in the germplasm is most important in plant improvement programmes for selection of parents. Alghamdi (2004) explained that the estimation of heritability and prediction of genetic advance becomes prejudiced when there is no information on the genotype by environment interactions. Allard and Bradshaw (1964) indicated that the best genotype is the one that has a consistent high performance over a wide range of environments. Genotype x environment interaction highly affects the phenotypic performance of any genotype. Therefore, it plays a significant role in the success of any breeding programme or the development of genetic material, adapted to several environments.

Beaver and Johnson (1981) stated that most soybean breeders give emphasis to wide adaptation rather than specific adaptation in their breeding programmes and select genotypes that perform well over a wide range of environments. Bekheit (2000) performed an experiment where 15 soybean genotypes were evaluated under three sowing dates during three seasons. The data established that low yielding genotypes tend to have high stability in different environments, and vice versa; high yielding genotypes were more likely to have lower stability. According to Gebeyehu and Assefa (2003), selection of genotypes based on the highest yielding, seemed less stable than the average of all genotypes, and selection exclusively for grain yield could result in disposing several stable genotypes. In soybean, yield differences of genotypes across environments and years has been linked with changes in number of seeds per unit area and this yield component is highly determined during a period from flowering to pod setting.

The performance of a crop in an environment is influenced by weather conditions, thus cultivars are vulnerable to environmental variation that in turn is a barrier to improving yield potential in the cultivars. Therefore, in any breeding programme the goal should be to create lines that are adapted to a wide range of environments. The other way would be to consider the test environment. It should at least include those representing yearly weather fluctuations as well as those imposed by different farmers’ practices. Alghamdi, (2004) stated that for soybean yield potential to meet the future demands, research should focus on understanding the physiological causes of genotype x environment interactions for genetic improvement.

1.7 Methods for analysing GEI and stability

There are many different ways of estimating phenotypic stability of genotypes. For example, Finlay and Wilkinson (1963) stated that the regression coefficient of varietal means on environmental means could be used as an indicator for phenotypic stability. Eberhart and Russell (1966) noted that regression techniques allow the genotype x environment interactions of each genotype to be partitioned into two parts; firstly, the portion of GEI due to the response in performance of the genotype to environments of varying levels of productivity, and secondly the portion due to deviations from regression.

Analysis of variance of multi-location trials is important for estimating variance components related to different sources of variation. These include genotypes and genotype by environment interaction (Crossa, 1990). The variance component analysis is crucial as it measures the errors that result from genotype by environment interaction in measuring traits such as yield. Hence, the knowledge of the magnitude of the interaction helps in estimating the genotypic effects and determining the optimum resource allocations in terms of number of sites and plots to be included in the next trial (Crossa, 1990). However, Romagosa *et al.* (1993) reported that when dealing with a large number of genotypes, estimating GEI using ANOVA is demanding and it fails to show the pattern of the GEI variance components.

Additive main effects and multiplicative interactions (AMMI) and genotype plus genotype by environment interaction (GGE) models effectively capture the additive (linear) and multiplicative (bilinear) components of GEI interaction and provides useful interpretation of multi-environment data in breeding programmes (Saini and Chetan, 2007). AMMI combines the additive components in a single model for the main effects of genotype, environment and multiplicative components for the interaction effect (Mitrovia *et al.*, 2012). Genotypic performances and phenotypic stability of the cultivars are best expressed by their graphic analyses (Miranda *et al.*, 2009) and it is useful in summarizing and emulating the response patterns that originally existed in the raw data. The GGE biplot analysis is another method that incorporates the genotype and GEI effects in the evaluation of cultivars and it uses graphic axes to identify best performing cultivars in the mega-environments (Akcura *et al.*, 2011). Secondly, this model provides genotype estimates in different locations. The other strength of this model is that it also combines ANOVA and PCA by separating sum of squares of genotypes and GEI together using the PCA method (Abay and Bjornstad, 2009). AMMI and

GGE statistical tools have huge importance and relevance to agricultural scientists because they deal with data that come from various types of experiments (Rad *et al.*, 2013).

1.8 Multi-trait relationships

GGE biplots have been used for analysing multi environment trials and the concept can be used to analyse data for multiple traits across locations. When performing multi-trait analyses, the genotypes are used as entries instead of using environments and traits are used as testers to construct genotype by trait (GT) biplots. This is an effective tool that graphically summarizes the genotype by trait data, visualises relationships among the measured traits and visualises the performance of genotypes based on the traits that influence selection of potential parents (Yan and Tinker, 2005). In addition, it helps identify less important traits that do not contribute directly to the trait of interest. Genotype by trait has been used in soybean yield analysis by Yan and Kang (2002) who reported that one genotype performed the best across all locations.

Most of the main breeding traits have negative correlations existing among them hence selection for a single trait is very difficult (Arshad *et al.*, 2004). Correlation coefficients are useful in quantifying the trait associations in terms of size and magnitude. However, they might be misleading if there is a large correlation that is a result of indirect effect of a trait. In soybean, scientists have used path analysis to partition correlations into direct and indirect traits (Haghi *et al.*, 2012).

1.9 Genetic diversity analysis in soybean

Genetic diversity analysis is of great importance to plant breeders as it helps selection of good parents for hybridisation for a successful breeding programme. Genetically diverse parents are useful to create variation for selection of useful recombinants with a high probability of high heterotic effects (Carpentieri-Pípolo *et al.*, 2003). Utilization of genetic diversity for any of the economically important traits present in landraces, cultivars and wild relatives aims at pyramiding of genes for better quality, higher productivity, and resistance to biotic and abiotic stresses (Dong *et al.*, 2001). This diversity is brought out through mutations or migration. In populations, genetic distances and number of alleles per locus among populations can be estimated using molecular markers (Nkongolo and Nsapato, 2003). Soybean, being a self-pollinated crop with limited outcrossing, has a narrow genetic base. However, it is believed that there is untapped diversity that is to be fully utilised for soybean improvement to broaden the genetic base (Dong *et al.*, 2004). Therefore, assessing the genetic diversity in soybean is the first step to achieving the goal of broad genetic diversity. This can only be achieved with accuracy by using molecular markers that are more reliable, stable and informative as compared to morphological diversity and pedigree analysis (Kumawat *et al.*, 2015).

1.10 Molecular marker characterisation

Since the late 19th century, plant breeders relied on phenotypic selection to improve plant varieties to achieve breeding progress through the assessment of external and internal traits such as disease resistance, yield, or quality traits (Bernardo, 2008). The selection of new, improved varieties that were developed was done by merely choosing genotypes with the desired phenotypes. The process of developing a new improved variety through phenotypic selection is time demanding and can take up to more than 10 years. Plant breeding using molecular techniques is becoming more popular and their role in genetic improvement of soybean germplasm is more important. Nevertheless, molecular technologies on themselves, can never replace conventional plant breeding research, but they will increase and improve the efficiency of plant breeding.

Molecular markers are DNA sequences with a precise defined nucleotide distribution and order, strictly specific for different organisms. Cost of development, reliability, level of polymorphism, informativeness and the number of samples to be used are some of the vital factors to consider when selecting markers for different applications (Sudarić *et al.*, 2010). In soybean breeding, molecular marker applications are currently focused in four main areas: germplasm characterization, marker-assisted selection (MAS), marker-assisted backcrossing and gene discovery. Marker-assisted selection is one of the applications that is used more readily than the usual techniques to screen single traits, such as resistance or restorer genes; insect resistance (Zhu *et al.*, 2003), nematode resistance (Meksem *et al.*, 2001; Kim *et al.*, 2010), and pathogen resistance (Shi *et al.*, 2009).

1.10.1 Types of molecular markers

In general, the ability to apply molecular markers to recognize the genomic position of a particular plant gene of interest has played a vital role in modernising the science of plant breeding and genetics. In soybean, amplified fragment length polymorphisms (AFLP), restriction fragment length polymorphism (RFLP), single nucleotide polymorphisms (SNP) and simple sequence repeats (SSR) have been used comprehensively to study genetic diversity and map genomic location of quantitative trait loci for many agronomic, physiological and seed composition traits. Shi *et al.* (2015) did a study based on the genomic DNA sequences of 27 soybean lines with known soybean cyst nematode (SCN) phenotypes, Kompetitive Allele Specific PCR (KASP) assays were developed for two single nucleotide polymorphisms (SNPs) from Glyma08g11490 for the selection of the Rhg4 resistance allele.

1.10.2 Single nucleotide polymorphism (SNP) as markers for genetic diversity studies

Single-nucleotide polymorphism (SNP) are referred to as the alterations in single DNA bases between homologous DNA fragments along with small deletions and insertions. Deulvot *et al*. (2010) defined the single nucleotide polymorphism (SNP) as the single DNA base differences between DNA fragments including insertions and deletion. Because the SNP represents nucleotide variation (for example sequence AC**G**TATA instead of AC**T**TATA), they are potentially useful as genetic markers as they are able to distinguish one haplotype from another. SNP markers have been proven to be the most abundant sources of DNA polymorphisms (Vignal *et al.*, 2002). With these properties, they can be easily used for genetic diversity studies, genetic and association mapping, and genome wide selection. SNP genotyping has been conducted in other studies such as maize (Yan *et al.*, 2009) and pea (Deulvot *et al.*, 2010). However, there is not much information on SNP markers in soybean. Therefore, it is important to assess the genetic diversity of tropical soybean lines using the SNP markers.

1.11 Conclusion

From the review, it can be concluded that soybean is one of the most important legume crops that contributes to both food and livestock feed. The review revealed that:

Although genotype by environment interaction has been extensively conducted in most crops soybean inclusive, there is still a need to conduct multi location testing of lines that have just been developed from breeding programmes to help in selection for adaptability and stability.

Conventional breeding alone has been used but it has shortfalls. Molecular markers could fill the gap, as they are able to detect variation at gene level.

There is little work on genetic diversity of soybean for future breeding programmes, hence the focus of this study.

REFERENCES

Abate, T., A.D. Alene, D. Bergvinson, B. Shiferaw, S. Silim, A. Orr. 2012. Tropical grain legumes in Africa and south Asia: knowledge and opportunities International Crops Research Institute for the Semi-Arid Tropics.

Abay, F. and A. Bjørnstad. 2009. Specific adaptation of barley varieties in different locations in Ethiopia. Euphytica 167: 181-195.

Akcura, M., S. Taner and Y. Kaya. 2011. Evaluation of bread wheat genotypes under irrigated multi-environment conditions using GGE biplot analyses. Agriculture 98: 35-40.

Alberts, M.J. 2004. A comparison of statistical methods to describe genotype x environment interaction and yield stability in multi-location maize trials. (Doctoral dissertation, University of the Free State).

Alghamdi, S.S. 2004. Yield stability of some soybean genotypes across diverse environments. Pakistan Journal of Biological Sciences 7: 2109-2114.

Allard, R.W. and A.D. Bradshaw. 1964. Implications of genotype-environmental interactions in applied plant breeding. Crop Science 4: 503-508.

Arshad, M., A. Bakhsh and A. Ghafoor. 2004. Path coefficient analysis in chickpea (*Cicer arietinum* L.) under rainfed conditions. Pakistan Journal of Botany 36: 75-82.

Baker, J., L. Allen, K. Boote, P. Jones and J. Jones. 1989. Response of soybean to air temperature and carbon dioxide concentration. Crop Science 29: 98-105.

Beaver, J. and R. Johnson. 1981. Yield stability of determinate and indeterminate soybeans adapted to the northern United States. Crop Science 21: 449-454.

Bekheit, M. 2000. Evaluation of some soybean genotypes in Upper Egypt. Thesis submitted to Faculty of Agriculture in partial fulfilment of requirement for the degree of M.Sc (Agriculture) in Assuit University., Egypt.

Bernardo, R. 2008. Molecular markers and selection for complex traits in plants: learning from the last 20 years. Crop Science 48: 1649-1664.

Borrás, L., G.A. Slafer and M.a.E. Otegui. 2004. Seed dry weight response to source–sink manipulations in wheat, maize and soybean: a quantitative reappraisal. Field Crops Research 86: 131-146.

Breene, W., S. Lin, L. Hardman and J. Orf. 1988. Protein and oil content of soybeans from different geographic locations. Journal of the American Oil Chemists' Society 65: 1927-1931.

Burton, J., R. Wilson, G. Rebetzke and V. Pantalone. 2006. Registration of N98–4445A mid-oleic soybean germplasm line. Crop Science 46: 1010-1012.

Burton, J.W. and L. Miranda. 2013. Soybean improvement: Achievements and challenges. Ratarstvo i Povrtarstvo 50: 44-51.

Carpentieri-Pípolo, V., F.A.M. da Silva and A.L. Seifert. 2003. Popcorn parental selection based on genetic divergence. Crop Breeding and Applied Biotechnology 3:90-106.

Catchot, A. 2010. Insect control guide for agronomic crops. Publication 2471: Extension Service of Mississippi state University, U.S.A.

Chianu, J., B. Vanlauwe, J. Mukalama, A. Adesina and N. Sanginga. 2006. Farmer evaluation of improved soybean varieties being screened in five locations in Kenya: Implications for research and development. African Journal of Agricultural Research 1: 143-150.

Cianzio, S., P. Arelli, B. Diers, H. Knapp, P. Lundeen, N. Rivera-Velez, et al. 2007. Soybean germplasm lines AR4SCN, AR5SCN, AR6SCN, AR7SCN, and AR8SCN. ISURF# Iowa State University, Ames, IA, USA.

Cooper, M., F. van Eeuwijk, S.C. Chapman, D.W. Podlich and C. Löffler. 2006. Genotype-by-environment interactions under water limited conditions. African Crop Science Journal 13: 41-47

Crossa, J. 1990. Statistical analyses of multilocation trials. Advances in Agronomy 44: 55-85.

Deulvot, C., H. Charrel, A. Marty, F. Jacquin, C. Donnadieu, I. Lejeune-Hénaut. 2010. Highly-multiplexed SNP genotyping for genetic mapping and germplasm diversity studies in pea. BMC Genomics 11: 468. doi:10.1186/1471-2164-11-468.

Dong, Y., L. Zhao, B. Liu, Z. Wang, Z. Jin and H. Sun. 2004. The genetic diversity of cultivated soybean grown in China. Theoretical and Applied Genetics 108: 931-936.

Dong, Y.S., B.C. Zhuang, L.M. Zhao, H. Sun and M.Y. He. 2001. The genetic diversity of annual wild soybeans grown in China. Theoretical and Applied Genetics 103: 98-103. doi:10.1007/s001220000522.

Dornbos, D. and R. Mullen. 1992. Soybean seed protein and oil contents and fatty acid composition adjustments by drought and temperature. Journal of the American Oil Chemists' Society 69: 228-231.

Eberhart, S.t. and W. Russell. 1966. Stability parameters for comparing varieties. Crop Science 6: 36-40.

Fabiyi, E. 2006. Soyabean processing, utilization and health benefits. Pakistan Journal of Nutrition 5: 453-457.

FAOSTAT 2016. Agriculture Organization of the United Nations Statistics Division (2014). Production Available in: http://faostat3. FAO. org/browse/Q/QC/S [Review date: April 2015].

Finlay, K. and G. Wilkinson. 1963. The analysis of adaptation in a plant-breeding programme. Australian Journal of Agricultural Research 14: 742-754.

Gandhi, A. 2009. Quality of soybean and its food products. International Food Research Journal 16: 11-19.

Gebeyehu, S. and H. Assefa. 2003. Genotype X environment interaction and stability analysis of seed yield in navy bean genotypes. African Crop Science Journal 11: 1-7.

Gibson, L. and R. Mullen. 1996. Soybean seed composition under high day and night growth temperatures. Journal of the American Oil Chemists' Society 73: 733-737.

Gurmu, F., H. Mohammed and G. Alemaw. 2009. Genotype x environment interactions and stability of soybean for grain yield and nutrition quality. African Crop Science Journal 17:87-98

Haghi, Y., P. Boroomandan, M. Moradin, M. Hassankhali, P. Farhadi, F. Farsaei. 2012. Correlation and path analysis for yield, oil and protein content of Soybean (*Glycine max* L.) genotypes under different levels of nitrogen starter and plant density. Biharean Biologist 6: 32-37.

IITA, 2009. Soybean. http://iita.org/web/iita-old/soybean. Accessed on 28 October 2018.

Kehlenbeck, H., C. Krone, E.-C. Oerke and F. Schönbeck. 1994. The effectiveness of induced resistance on yield of mildewed barley Journal of Plant Diseases and Protection 101: 11-21.

Kim, M., D.L. Hyten, A.F. Bent and B.W. Diers. 2010. Fine mapping of the SCN resistance locus from PI 88788. The Plant Genome 3: 81-89.

Krisnawati, A. and M.M. Adie. 2015. Selection of Soybean Genotypes by Seed Size and its Prospects for Industrial Raw Material in Indonesia. Procedia Food Science 3: 355-363. doi:http://dx.doi.org/10.1016/j.profoo.2015.01.039.

Kumawat, G., G. Singh, C. Gireesh, M. Shivakumar, M. Arya, D.K. Agarwal. 2015. Molecular characterization and genetic diversity analysis of soybean (*Glycine max* (L.) Merr.) germplasm accessions in India. Physiology and Molecular Biology of Plants 21: 101-107. doi:10.1007/s12298-014-0266-y.

Kumudini, S. 2010. Soybean growth and development. The Soybean: Botany, Production and Uses. British Library, London, UK: pp48-73.

Meksem, K., P. Pantazopoulos, V. Njiti, L. Hyten, P. Arelli and D. Lightfoot. 2001. 'Forrest'resistance to the soybean cyst nematode is bigenic: saturation mapping of the Rhg1and Rhg4 loci. Theoretical and Applied Genetics 103: 710-717.

Miladinović, J., M. Vidić, V. Đorđević and S. Balešević-Tubić. 2015. New trends in plant breeding-example of soybean. Genetika 47: 131-142.

Miranda, G.V., L.V.d. Souza, L.J.M. Guimarães, H. Namorato, L.R. Oliveira and M.O. Soares. 2009. Multivariate analyses of genotype x environment interaction of popcorn. Pesquisa Agropecuária Brasileira 44: 45-50.

Misiko, M., P. Tittonell, J. Ramisch, P. Richards and K. Giller. 2008. Integrating new soybean varieties for soil fertility management in smallholder systems through participatory research: Lessons from western Kenya. Agricultural Systems 97: 1-12.

Mitroviã, b., s. Treski, m. Stojakoviã, m. Ivanoviã and G. Bekavac. 2012. Evaluation of Experımental Maize Hybrids Tested in Multi-Location Trials Using AMMI and GGE Biplot Analyses. Turkish Journal of Field Crops 17: 35-40.

Nkongolo, K. and L. Nsapato. 2003. Genetic diversity in *Sorghum bicolor* (L.) Moench accessions from different ecogeographical regions in Malawi assessed with RAPDs. Genetic Resources and Crop Evolution 50: 149-156.

Pathan, M. and D.A. Sleper. 2008. Advances in soybean breeding. Genetics and genomics of soybean. Springer p. 113-133.

Rad, M.N., M.A. Kadir, M. Rafii, H.Z. Jaafar, M. Naghavi and F. Ahmadi. 2013. Genotype environment interaction by AMMI and GGE biplot analysis in three consecutive generations of wheat (*Triticum aestivum*) under normal and drought stress conditions. Australian Journal of Crop Science 7: 956.

Reinprecht, Y., S.-Y. Luk-Labey, K. Yu, V.W. Poysa, I. Rajcan, G.R. Ablett. 2011. Molecular basis of seed lipoxygenase null traits in soybean line OX948. Theoretical and Applied Genetics 122: 1247-1264.

Romagosa, I., P. Fox, L.G. Del Moral, J. Ramos, B.G. del Moral, F.R. De Togores. 1993. Integration of statistical and physiological analyses of adaptation of near-isogenic barley lines. Theoretical and Applied Genetics 86: 822-826.

Saini, P. and S. Chetan, 2007. A Review on Genotype Environment Interaction and its Stability Measures. Plant Breeding Reviews 15:155-204

Shi, A., P. Chen, D. Li, C. Zheng, B. Zhang and A. Hou. 2009. Pyramiding multiple genes for resistance to soybean mosaic virus in soybean using molecular markers. Molecular Breeding 23: 113-124.

Shi, Z., S. Liu, J. Noe, P. Arelli, K. Meksem and Z. Li. 2015. SNP identification and marker assay development for high-throughput selection of soybean cyst nematode resistance. BMC Genomics 16: 314-321.

Singh, B. (1991). Tropical grain legumes as important human foods. Economic Botany 46(3): 310-321.

Sionit, N., B. Strain and E. Flint. 1987. Interaction of temperature and CO2 enrichment on soybean: growth and dry matter partitioning. Canadian Journal of Plant Science 67: 59-67.

Sudarić, A., M. Vratarić, S. Mladenović-Drinić and M. Matosa. 2010. Biotechnology in soybean breeding. Genetika 42: 91-102.

Tefera, H. 2011. Breeding for promiscuous soybeans at IITA. Soybean-Molecular Aspects of Breeding. InTech.

Tefera, H., A. Kamara, B. Asafo-Adjei and K. Dashiell. 2009. Improvement in grain and fodder yields of early-maturing promiscuous soybean varieties in the Guinea Savanna of Nigeria. Crop Science 49: 2037-2042.

Vignal, A., D. Milan, M. SanCristobal and A. Eggen. 2002. A review on SNP and other types of molecular markers and their use in animal genetics. Genetics Selection Evolution 34: 275-281.

Yan, J., T. Shah, M.L. Warburton, E.S. Buckler, M.D. McMullen and J. Crouch. 2009. Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. PloS one 4: e8451.

Yan, W. and L. Hunt. 1998. Genotype by environment interaction and crop yield. Plant Breeding Reviews 16: 135-178.

Yan, W. and M.S. Kang. 2002. GGE biplot analysis: A graphical tool for breeders, geneticists, and agronomists. CRC press.

Yan, W. and N.A. Tinker. 2005. An integrated biplot analysis system for displaying, interpreting, and exploring genotype× environment interaction. Crop Science 45: 1004-1016.

Zhu, Y., Q. Song, D. Hyten, C. Van Tassell, L. Matukumalli, D. Grimm. 2003. Single-nucleotide polymorphisms in soybean. Genetics 163: 1123-1134.

CHAPTER 2

YIELD STABILITY AND ADAPTATION ANALYSIS OF ELITE SOYBEAN LINES ACROSS DIVERSE ENVIRONMENTS IN SOUTHERN AFRICA

ABSTRACT

Soybean is an important legume crop that is a source of essential amino acids for human consumption and livestock. The multi-location testing of soybean genotypes precedes selection in plant breeding programmes. The study was carried out to investigate genotype by environment interaction (GEI) for grain yield in 26 elite soybean lines along with four checks in six environments spreading over three countries (Malawi, Mozambique and Zambia) in a 6 x 5 alpha lattice design. The additive main effect and multiplicative interaction model (AMMI) indicated that environment, genotypes and genotype by environment interaction significantly affected grain yield ($P<0.001$) and contributed 3.8%, 17% and 78%, respectively, to the total variation. Three AMMI interaction principal components (IPCA1, IPCA2 and IPCA3) were significant ($P<0.01$). The genotype, genotype x environment interaction (GGE) biplots were created based on the first two principal components PC1 and PC2 that accounted for 39.23 and 26.86% of genotype plus GEI variation, respectively. The GGE biplot analysis ranked the genotypes for yield and stability, and environments for representativeness and discriminativeness. The relationships between genotypes and environments were also demonstrated. Genotype G4 (TGX 2001-3FM) was identified as the ideal genotype with high grain yield mean performance and high stability. Therefore, it could be recommended for cultivar release if the study can be repeated to verify these findings. The environment E6 (Nampula, Mozambique) was the most informative test environment, hence it is ideal for selecting generally adapted genotypes. Genotypes G11 (TGX 2002-4FM) and G22 (TGX 2001-15DM) were low yielding but with high stability hence can be recommended for further yield improvement.

Key words: AMMI, GGE biplot, G X E interaction, Soybean [*Glycine max* (L.) Merrill]

2.1 Introduction

Soybean (*Glycine max* (L.) Merrill) is one of the world's leading legume crops that is a source of oil (20%) and protein (40%). It is used for human diet, animal feed, improving soil fertility and as a raw material in several manufacturing industries. Soybean demand in Africa is more than the supply, as such, its production has increased and this trend is growing exponentially. In Africa, South Africa is the leading producer contributing 35% of the total production seconded by Nigeria (27%) and Uganda (8.5%) (Murithi *et al.*, 2015). Currently, Zambia, Malawi and Zimbabwe still contribute a significant amount of 1.5 million tonnes in total to production in sub Saharan Africa (Abate *et al.*, 2012). Soybean production in Zambia, Malawi and Zimbabwe is still projected to grow to 2 million tonnes by 2020 to meet the demand.

The soybean crop is grown in areas that have different climatic conditions such as temperature, rainfall and soil characteristics (Branquinho *et al.*, 2014). As a result, the cultivars perform differently in the different environments resulting in genotype x environment interaction (GEI). This concept of GEI is common in multi environment yield trials making selection of superior, stable cultivars in a growing region very difficult. The limitations presented on variety selection can be evaded through selection of genotypes for a specific environment or widely adapted and stable genotypes across environments (Ceccarelli, 1989).

Yield stability of genotypes across environments is an important phenomenon in plant breeding, as it helps to make recommendations as to whether that genotype is best for wide or specific production in the environments. A stable genotype is the one that has the ability of using all the resources in high yielding environments and has its overall mean performance above average in all environments (Allard and Bradshaw, 1964). Eberhart and Russell (1966) added that, the candidate genotypes, that are ideal for stability testing, should have the genetic potential for best performance under the targeted environments. Plant breeders mostly use genotype by environment interaction stability statistics to assess the performance of their genotypes across various environments established from the information acquired from the evaluation of cultivars grown in a sample of growing environments.

Several statistical methodologies have been used to evaluate and analyse the performance of soybean lines for selecting the most stable and productive line(s) for locations and regions. The additive main effect and multiplicative interaction (AMMI) analysis is a statistical tool used to evaluate the effects of GEI (Gauch, 2006). In addition, the genotype plus genotype x environment interaction (GGE) models proposed by Yan and Kang (2002) have been emphasized for multi environment trial data. Several researchers have used these models in their soybean studies (Oliveira *et al.*, 2016; Atnaf *et al.*, 2013). GGE biplots best fit for mega-environment analysis involve the ‘which-won-where’ pattern; for genotype evaluation, the mean vs. stability; and for test environment evaluation, for discriminating power vs. representativeness of the test environments (Yan and Kang, 2002). This tool estimates the effects of genotypes along with the GEI and it has been known as a useful method to visualize and analyse the pattern of GEI in multi environment evaluation of different crops such as soybean, maize, wheat and oilseeds (Asfaw *et al.*, 2009; Mohammadi *et al.*, 2010; Nzuve *et al.*, 2013).

Soybean yield and other agronomic traits are strongly affected by GEI (Alghamdi, 2004). Individual genotypes of soybean are adapted to regions differently since their phenotypes are highly influenced by the genotype and environments that they are grown in. This has also been observed in other studies including maize (Sibiya *et al.*, 2012), wheat (Mohammadi *et al.*, 2017) and cassava (Chipeta *et al.*, 2017). Therefore, there is a need to evaluate soybean genotypes in order to understand and visualise their performance across the growing locations in some southern parts of Africa. The objective of this study was, therefore, to determine the adaptability and stability of grain yield of elite soybean lines using AMMI and GGE biplot analyses methods.

2.2 Materials and methods

2.2.1 Genotypes and evaluation environments

Twenty-six elite soybean lines along with four checks were used to generate data used in this analysis during the 2016/2017 growing season in three countries (Malawi, Mozambique and Zambia) and six locations. The list of genotypes used and the geographical information of each location (environment) are shown in Tables 2.1 and 2.2, respectively. Of the 30 genotypes evaluated, one (TGX 2002-3DM) was early maturing and determinate in growth habit while the rest were medium maturing and indeterminate in growth habit.

2.2.2 Design of trials and agronomic management

The soybean genotypes were planted in a 6 x 5 alpha lattice design with three replications. Each genotype occupied a plot comprising of four rows of 4 m in length, 0.5 m between rows and 0.05 m intra-row spacing. At maturity, grain yield was estimated from the two middle rows, which was considered as a net plot, leaving one row at 0.5 m on either sides as borders. Manual and chemical weeding was done to mitigate weeds and inorganic fertilizers were applied at planting at all locations depending on the results from soil analysis.

2.2.3 Data collection and analysis

Grain yield data was collected for each genotype at all evaluation environments. The data were subjected to analysis of variance (ANOVA) across locations and at each location, AMMI ANOVA across locations, and GGE biplot analysis.

The model (equation 2.1) for the combined ANOVA of multi-environment trials was used in GenStat 17th edition (Payne et al., 2014). The model includes additive terms for main effects of genotype and environment, as well as the genotype by environment interaction term.

$$\mathrm{Y}_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \mathrm{E}_{ij} \quad \text{.. Equation 2.1}$$

Where Yij is the yield of the genotype i in environment j and *k*th replication;, μ is overall yield mean, αi and βj are genotypic and environmental effect, (αβ)ij is the effect of interaction between the ith genotype and jth environment, ∈ij is the mean random error of the ith genotype and ej environment.

The AMMI model used was adopted from Gauch and Zobel (1989) using the model in Equation 2.2 below

$$\mathrm{Y}_{ij} = \mu + \alpha_i + \beta_j + \sum_n \lambda_n \delta_{in} \gamma_{jn} + \mathrm{P}_{ij} + \mathrm{E}_{ij} \quad \text{...Equation 2.3}$$

In this model, the *i*th is the genotype effect in *j*th environment and *k*th replication; and the additive components of the model which are μ is the grand mean, the *i*th genotype effect (αi), and the *j*th environment effect (βj). The terms λn δin γjn and p_{ij} constitute the multiplicative component, where λn, is the interaction principal component, αin, is the eigen vector for the genotypic principal component, γjn is the environmental principal component. Only the first or second interaction principal components (IPCAs) are retained for analysis and the rest of the interaction variation is explained by the residual p_{ij}. The last component in the model is ϵij, which is the random error. The contribution of each interaction principal component to the total genotype x environment interaction sum of squares was determined.

Biplots were plotted using the first two IPCAs to depict the relative performance of genotypes for yielding and stability.

The GGE model in Equation 2.3 below was used as adopted from (Yan and Kang, 2002)

$$\mathrm{Y}_{ij} - \mu - \beta_j = \lambda_1 \gamma_{i1} \delta_{j1} + \lambda_2 \gamma_{i2} \delta_{j2} + \mathrm{E}_{ij} \quad \text{.....................................Equation 2.4}$$

GGE biplots were constructed using least squares means for grain yield from each environment.

Where: Y_{ij} is the mean of *i*th genotype in the *j*th environment, μ is the grand mean, βj is the *j*th environment main effect, and μ +β_j is the mean of all genotypes in *j*th environment. The terms λ_1 and λ_2 are the singular values for the first (PC1) and second (PC2) principal components, respectively; γ_{i1} and γ_{i2} are eigenvectors of the *i*th genotype for PC1 and PC2, respectively. The components δ_{j1} and δ_{j2} are eigenvectors of the *j*th environment for the principal components PC1 and PC2, respectively; and $\in_{ij}$ is the residual associated with the *i*th genotype in the *j*th environment.

Biplots were plotted for comparing genotypes in regard to mean performance and stability across environments, compare environments to determine if there is similarities or differences in expression of grain yield among them, determine which genotypes were best yielders in particular environments, and to determine which genotypes and environments were best for expression of particular traits.

Table 2.1 Genotypes evaluated and their characteristics

Genotype	Genotype code	Source	Maturity	Growth habit
TGx 2001-12FM	G1	IITA	M	I
TGx 2001-16FM	G2	IITA	M	I
TGx 2001-24DM	G3	IITA	M	I
TGx 2001-3FM	G4	IITA	M	I
TGx 2001-4FM	G5	IITA	M	I
TGx 2001-5FM	G6	IITA	M	I
TGx 2002-12FM	G7	IITA	M	I
TGx 2002-1FM	G8	IITA	M	I
TGx 2002-3DM	G9	IITA	E	D
TGx 2002-3FM	G10	IITA	M	I
TGx 2002-4FM	G11	IITA	M	I
TGx 2002-9FM	G12	IITA	M	I
TGx 2001-19FM	G13	IITA	M	I
TGx 2001-7FM	G14	IITA	M	I
TGx 2002-10FM	G15	IITA	M	I
TGx 2001-14FM	G16	IITA	M	I
TGx 2001-27DM	G17	IITA	M	I
TGx 2001-6DM	G18	IITA	M	I
TGx 2002-6DM	G19	IITA	M	I
TGx 2001-13FM	G20	IITA	M	I
TGx 2001-14DM	G21	IITA	M	I
TGx 2001-15DM	G22	IITA	M	I
TGx 2001-21DM	G23	IITA	M	I
TGx 2001-21FM	G24	IITA	M	I
TGx 2002-11FM	G25	IITA	M	I
TGx 2001-1FM	G26	IITA	M	I
TGx 2001-26DM	G27	IITA	M	I
Tikolore	C1	IITA	M	I
NASOKO	C2	IITA	M	I
SC SERENADI	C3	IITA	M	I

IITA = International institute for tropical agriculture, M = medium, E = early, I = intermediate, D = Determinant

Table 2.2 Trial locations and their geographical information

Country	Location And Environment Code	Altitude	Latitude	Longitude	Min Temp	Max Temp
		m	^{0}S	^{0}E	^{0}C	^{0}C
Malawi	Chitedze E6	1149	13.9815	33.6372	18	29
	Bvumbwe E2	1146	15.917	35.067	14	25
Zambia	IITA SARAH E4	1199	15.302	27.574	15	30
	Kabwe E5	1182	14.4285	28.4514	17	28
Mozambique	Nampula E1	360	15.1266	39.2687	19	34
	Gurue E3	788	15.4914	37.0125	19	29

2.3 Results

2.3.1 Combined analysis of variance

The analysis of variance for grain yield showed significant differences (P<0.001) for the genotype, environment and the interaction between genotype and environment. The combined analysis of variance (Table 2.3) revealed that grain yield was affected significantly by environment, genotype and genotype by environment interaction that explained 3.2%, 10%, and 55% of the total variation, respectively. The coefficient of variation was 17.9% for the combined analysis for all trials in 6 sites with a grand mean yield of 1699 kg/ha.

Table 2.3 Analysis of variance for grain yield across six environments

Source	DF	SS	MS
SITE	5	5907786.94	1181557.39***
REP(SITE)	12	977691.47	81474.29ns
BLOCK(SITE*REP)	72	6302792.6	87538.79ns
ENTRY	29	19591813.02	675579.76***
SITE*ENTRY	145	96890342.83	668209.26***
Error	276	25709349.9	93149.8
Total	539	184498654.1	
Yield mean	1699 kg/ha		
CV	17.90%		

***Significant at P<0.001, ns = not significant, DF = Degrees of freedom, CV = Coefficient of variation, SS = Sum of squares, MS = Mean square

The soybean genotypes had different mean perfomances across the locations. Environments E4 and E6 had the highest mean perfomance and E1 had the lowest mean perfomance (Table 2.4). Genotypes G10 and G24 had the lowest mean performance across environments while G4, G13 and G22 had the highest perfomance across environments.

Table 2.4 Genotype mean perfomance across all six locations

Genotype code	E1	E2	E3	E4	E5	E6	Mean yield (t/ha)
G1	1.25	1.37	0.98	2.06	1.33	1.97	1.49
G2	2.44	2.07	1.90	1.64	2.45	1.59	2.01
G3	0.62	1.43	2.25	2.22	1.54	1.34	1.57
G4	1.67	2.53	1.34	1.29	2.64	3.04	2.08
G5	1.51	1.47	1.24	2.05	2.20	2.15	1.77
G6	1.14	2.74	1.39	2.82	1.19	1.44	1.79
G7	1.63	1.11	1.55	1.51	1.72	1.38	1.48
G8	2.00	2.62	1.57	2.45	1.39	1.65	1.95
G9	1.73	1.63	1.60	2.34	1.25	2.06	1.77
G10	0.73	0.86	2.13	1.28	1.49	1.43	1.32
G11	2.05	0.89	1.94	0.94	2.65	2.12	1.76
G12	1.76	1.64	1.32	0.78	1.13	2.47	1.52
G13	2.12	1.33	1.89	2.89	2.02	2.08	2.06
G14	1.11	1.26	2.38	2.18	1.45	1.41	1.63
G15	1.23	1.72	1.19	1.30	1.76	1.73	1.49
G16	2.14	2.18	1.46	2.03	1.74	2.13	1.95
G17	2.54	1.83	0.88	2.67	1.19	1.55	1.78
G18	1.42	1.52	2.52	2.72	0.85	1.68	1.78
G19	1.50	1.56	1.64	2.20	1.35	1.58	1.64
G20	0.47	1.32	2.15	2.44	2.12	2.46	1.83
G21	1.55	1.40	1.82	2.21	1.22	2.77	1.83
G22	2.02	1.33	3.22	1.63	2.04	2.13	2.06
G23	1.66	1.86	0.91	1.25	1.24	1.75	1.45
G24	1.44	1.41	1.21	1.42	1.35	0.99	1.30
G25	2.23	1.16	1.14	1.61	1.62	1.20	1.49
G26	1.52	1.81	2.05	1.83	2.18	1.52	1.82
G27	1.43	2.56	2.11	1.69	1.66	1.57	1.84
G28	1.14	1.78	2.43	1.48	1.91	1.37	1.68
G29	0.73	1.19	0.99	1.45	1.89	2.32	1.43
G30	1.13	2.73	1.46	0.93	0.84	1.43	1.42
Mean Yield (t/ha)	1.53	1.68	1.69	1.84	1.65	1.81	1.70

2.3.2 AMMI analysis

The additive main effect and multiplicative interaction (AMMI) analysis of variance showed significant effects for all genotypes, environment and the genotype by environment interaction (GEI) (Table 2.5). The partitioning of the variance components also showed that 3.8% of the total variation was contributed by the environment, 17.5% was due to the genotypes, and GEI were associated with 78% of the total variation, respectively. The three interaction principal components (IPCA1, IPCA2 and IPCA3) were significant (p<0.001). These IPCA's contributed 29.20%, 28.43% and 17.47%, respectively to the total interaction sum of squares and cumulatively they contributed 75.10% of the total genotype by environment interaction sum of squares.

Table 2.5 AMMI analysis for grain yield from across six locations

Source of variations	DF	SS	MS	Total variation (%)	GE explained (%)	GE cumulative (%)
Total	539	184506352	342312		-	-
Block (Env)	12	978137	81511			
Treatments	179	151518215	846470***		-	-
Genotypes	29	26626275	918147***	17.50	-	-
Environment	5	5907680	1181536***	3.80	-	-
GE	145	118984260	820581***	78.00		
IPCA 1	33	34775458	1053802***	-	29.20	29.20
IPCA 2	31	33832432	1091369***	-	28.43	57.63
IPCA 3	29	20783117	716659***		17.47	75.10
Residuals	52	29593254	569101***		24.87	-
Error	348	32010000	91983		-	-

*** Significant at P<0.001, DF = Degrees of freedom, SS = Sum of squares, MS = Mean sum of squares, GE = Genotype x Environment interaction, IPCA 1 = Interaction principal component axis 1, IPCA 2 = Interaction principal component axis 2, IPCA 3 = Interaction principal component 3

2.3.3 Mean grain yield and IPCA scores of genotypes

For the 30 genotypes evaluated, the mean yield ranged from 1.30 t/ha to 2.08 t/ha (Table 2.4). Among these genotypes, G24 had the lowest mean yield and cultivar G4 had the highest mean yield. Fifty-three percent of the genotypes in the study (G2, G4, G5, G6, G8, G9, G11, G13, G16, G17 G20, G21, G22, G18, G26 and G27) performed above the grand mean of 1.7 t/ha.

Table 2.6 Mean yield first, second and third IPCA scores of genotypes

Genotype code	Mean GY (t/ha)	IPCAg[1]	IPCAg[2]	IPCAg[3]
G1	1.49	-0.07	-0.21	0.21
G2	2.01	0.11	-0.08	-0.61
G3	1.57	0.49	0.04	0.11
G4	2.08	-0.69	0.15	-0.25
G5	1.77	-0.07	-0.02	-0.05
G6	1.79	0.19	-0.41	0.25
G7	1.48	0.21	0.06	-0.28
G8	1.95	0.07	-0.58	0.08
G9	1.77	0.07	-0.23	0.21
G10	1.32	0.3	0.6	-0.05
G11	1.76	-0.01	0.53	-0.71
G12	1.52	-0.54	0.04	-0.22
G13	2.06	0.37	-0.17	0.06
G14	1.63	0.49	0.19	0.27
G15	1.49	-0.17	0	-0.16
G16	1.95	-0.14	-0.32	-0.15
G17	1.78	0.16	-0.86	-0.12
G18	1.78	0.51	-0.01	0.43
G19	1.64	0.21	-0.2	0.15
G20	1.83	0.13	0.46	0.6
G21	1.83	-0.14	0.11	0.52
G22	2.06	0.33	0.61	-0.21
G23	1.45	-0.29	-0.31	-0.22
G24	1.30	-0.37	-0.01	0.05
G25	1.49	0.22	-0.29	-0.52
G26	1.82	0.23	0.11	-0.17
G27	1.84	-0.46	0.15	0.26
C1	1.68	0.24	0.35	-0.06
C2	1.43	-0.52	0.33	0.27
C3	1.42	-0.85	-0.01	0.29

The mean yield for the environments (Table 2.7) ranged from 1.5 t/ha to 2.1 t/ha. Two of the environments (E4 and E6) had their performance above the grand mean of 1.7t/ha. E1 had the lowest mean yield and E4 was recorded as the highest yielding environment.

Table 2.7 Mean yield first, second and third IPCA scores of environments

Environment	Code	Mean GY (t/ha)	IPCAe[1]	IPCAe[2]	IPCAe[3]
Nampula	E1	1.5	0.1	-0.72	-1.09
Bvumbwe	E2	1.7	-0.71	-0.66	0.15
Gurue	E3	1.7	0.81	1.03	0.2
IITASARAH	E4	1.8	0.92	-0.77	0.91
Kabwe	E5	1.7	0.17	0.65	-0.69
Chitedze	E6	2.0	-1.29	0.48	0.52

2.3.4 Best four selections per environment

The best four genotypes per environment were identified using the AMMI analysis (Table 2.8). G4 was the best in one environment and ranked second and third in two other environments followed by G22, which was also best in one environment and ranked third and fourth in two other environments. Genotypes G11, G17, G18, and G30 performed best in one environment each, and G11 was second in another environment.

Table 2.8 First four selections per environment

		Ranking per environment			
Environment	Mean (t/ha)	1	2	3	4
E5	1.7	G11	G4	G22	G2
E3	1.7	G22	G20	G3	C1
E6	2.0	G4	G11	G2	G22
E1	1.5	G17	G13	G4	G16
E4	1.8	G18	G13	G17	G6
E2	1.7	C3	G6	G8	G27

2.3.4 AMMI biplot: IPCA1 *vs* IPCA2

The AMMI biplot analysis (Figure 2.1) revealed that environment E6, E3 and E4 had the longest vectors compared to E2, E5 and E1. Genotype G17 had specific adaptation with high yielding environments. Cultivars G12, G4, G27, G29 and G21 had a positive interaction with environment E6, hence were specifically adapted to E6. The following cultivars; G2, G5, G15, and G21 were all close to the centre of the biplot. Two sets of environments (E1 and E4, plus E3 and E5) had acute angles in between them. The biplot analysis of GE based on the AMMI2 model for the first two interaction principal component scores, namely IPCA1 and IPCA2, revealed that the two IPCAs cumulatively contributed 66.09% of the GE.

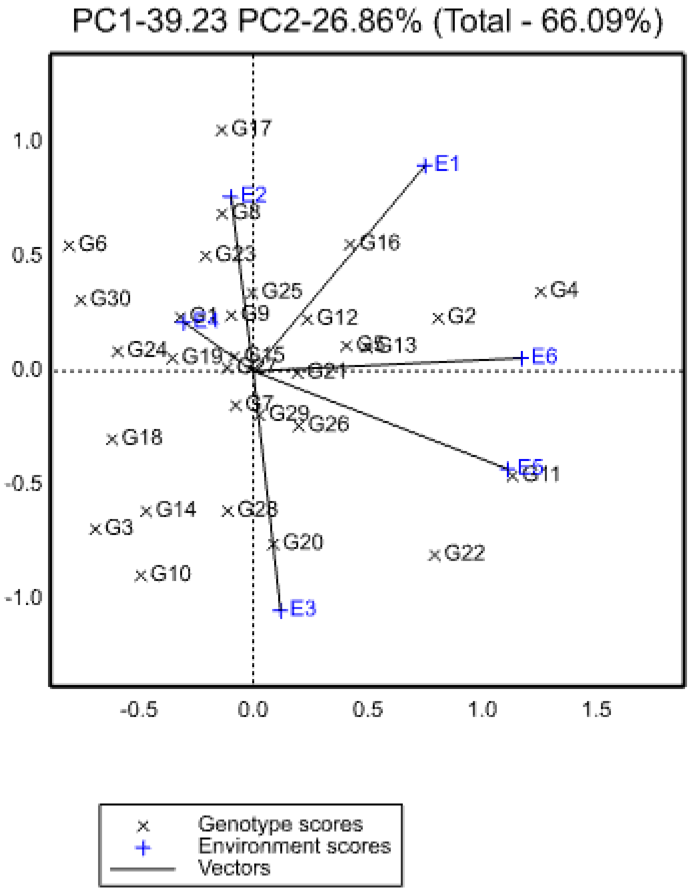


Figure 2.1 AMMI biplot analysis based on the first two interaction principal components

2.3.6 GGE Analysis

Results of the GGE biplots as presented in figs 2.2-2.6, shows that the first and second principal components explained a total variation of 65.71%.

2.3.6.1 Relationship among environments

The lines drawn from the origin of the biplot connecting to the environment markers are called environment vectors (Figure 2.2). The angles between these vectors indicate the correlation between the environments. Angles between vectors that are less than 90^0 shows that there is a high correlation between the environments as observed among E3 and E2, E6 and E5, E1 and E6, E5 and E3. Vectors for E4 and E1 were at right angles and the angle between E4 and E5 and E3 was more than 90^0. Both the angle and length of the environment vectors indicate the similarity between the test environments. Among the test locations E1, E6, E5 and E3 had the longest vectors and environments E2 and E4 had shorter vectors. Genotypes such as G15 and G27 were clustered closer to the point of origin of the biplot. (Figure 2.2). G4 had the highest mean followed by G2, G11 and G16.

2.3.6.2 "Which-won-where" polygon view

The polygon was formed by connecting the genotypes G4, G10, G22, G3, G17 and G6 farthest from the point of origin of the biplot (Figure 2.3). The genotypes on the vertices performed either the best or poorly in one or more of the test environments. The highest performing genotype in environments E6 and E1 was G4; while in E5 and E2, the best performing genotypes were G11 and G17 respectively. C3 and G6 were the highest performing in environment E4. However, genotypes G8, G18 and G10 performed poorly in the test environments. There were seven rays that divided the biplot into 7 sectors. Rays are the perpendicular lines to the sides of the polygon formed in the plot (Kaya *et al.,* 2006). The environments fell into different sectors except for E1 and E6, which fell into the same sector.

2.3.6.3 Genotype and environment comparisons

The grain yield performance of genotypes (Figure 2.4) showed that G2, G4 and G12 had the highest mean yield. E6 was in the first concentric circle closer to the centre of the ideal environment (Figure 2.5) and E3 was the furthest from the centre of the concentric circles.

2.3.6.4 Mean versus stability

Genotypes such as G2, G24 G19 and G4 had short vectors running from the AEC while genotypes G22, G11, G10, G17 and G6 had the longest vectors. The AEC ordinate divided the genotypes into two groups those above it; from G4 to G17 had high mean performance and those below it, from G7 to G3 had low mean performance (Figure 2.6)

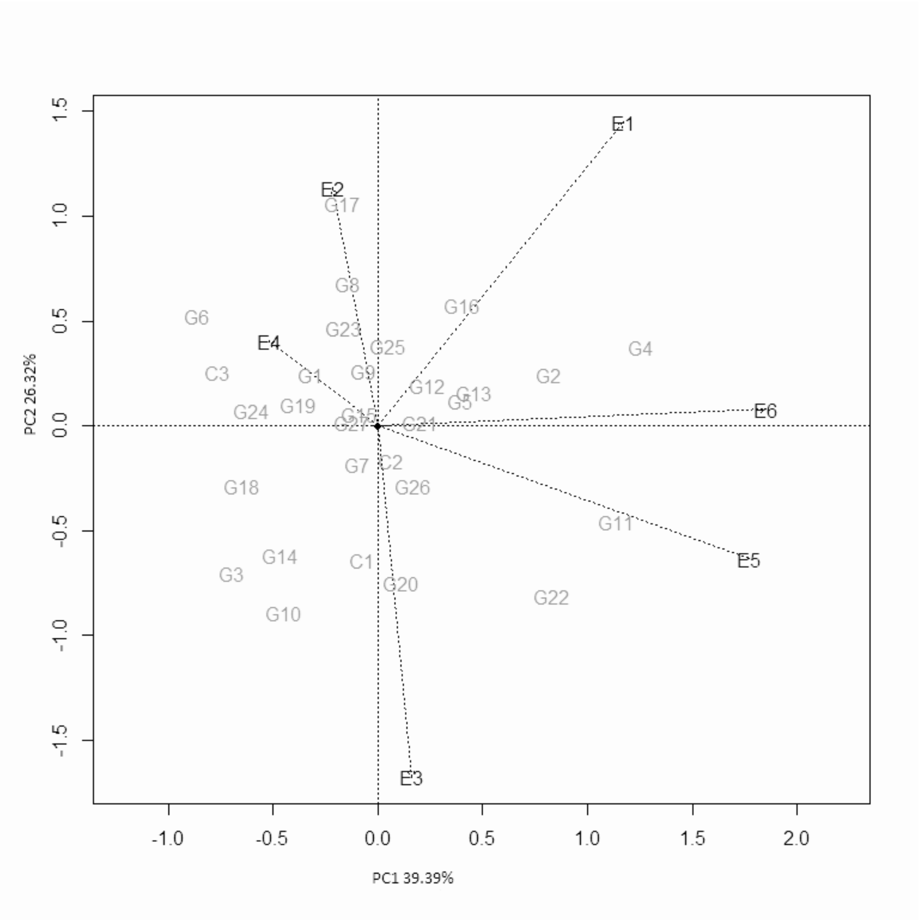


Figure 2.2 Environment vector view to show relationships among test environments

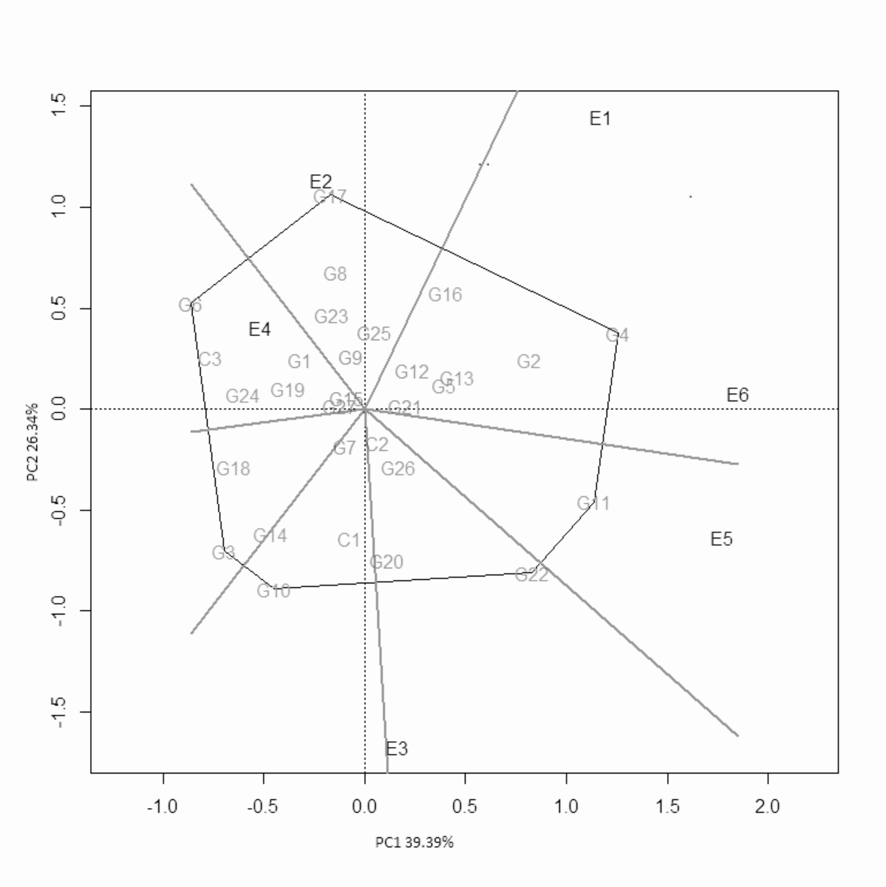


Figure 2.3 The "which won where" view of the GGE biplot

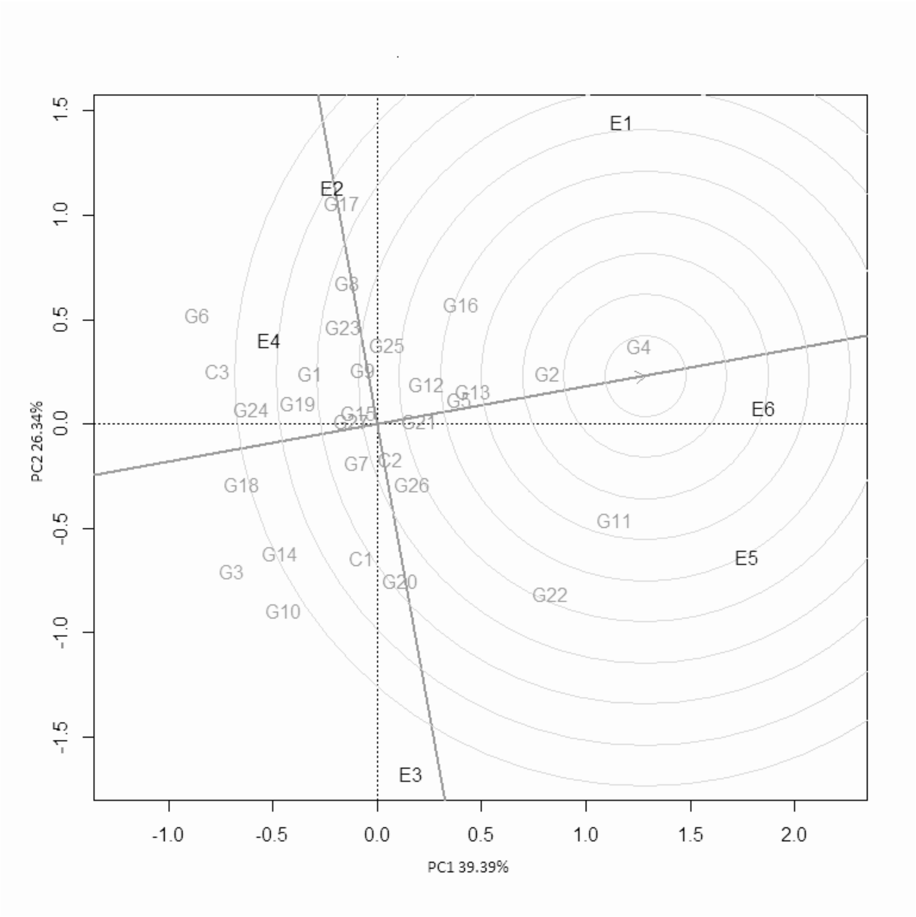


Figure 2.4 Genotype comparison with the ideal genotype

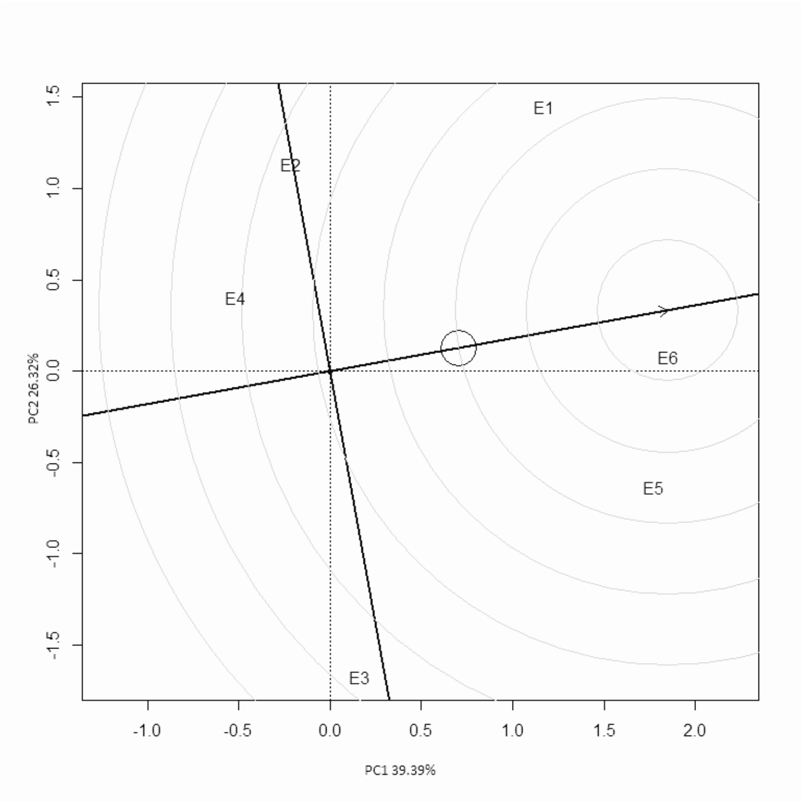


Figure 2.5 Environment ranking and comparison with the ideal environment

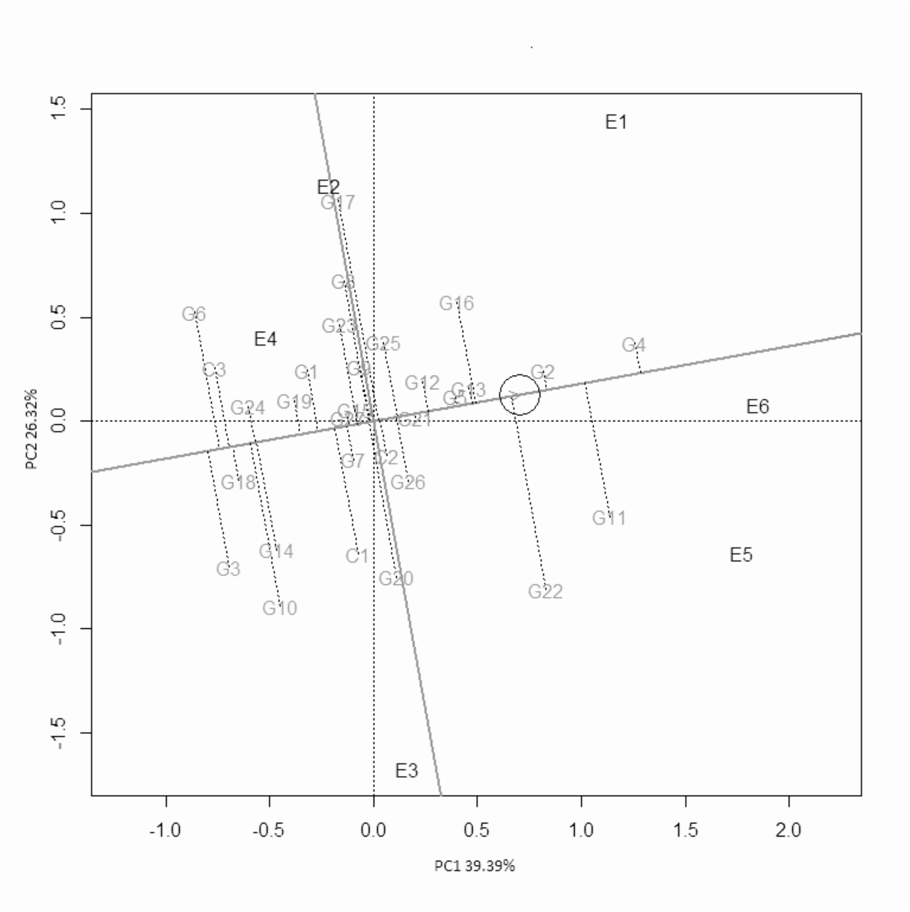


Figure 2.6 Ranking of genotypes based on mean perfomance and stability

2.4 Discussion

The study showed significant main effects for genotypes and environments indicating variation among the genotypes and test environments. The genotype × environment interaction (GEI) was also significant indicating differential ranking of genotypes across the environments. This GEI may confound the process of selecting superior genotypes, recommendation of a genotype for a target environment and reduce the selection efficiency in different breeding programmes (Gauch, 2006).

The ANOVA for grain yield using the AMMI method showed that environments (E), genotypes (G) and genotype × environment (GEI) interaction significantly affected the soybean grain yield. As per AMMI analysis, environment and genotype accounted for 17.5% and 3.8% of the total variation, respectively. The significant G×E interaction explained 78% of the variation that was almost triple that of the genotypic effects and ten times more of the environmental effect. Genotype and genotype by environment interaction are relevant to cultivar evaluation, especially when G×E interaction is determined as repeatable (Cooper and Hammer, 1996).

This also agrees with the findings of Bhartiya *et al.* (2017) who indicated that the GEI explained more variation compared to genotypes and environments. The performance of the soybean was different at different locations hence the large GEI effects realised in this study (Atnaf *et al.*, 2013).

Romagosa *et al.* (1993) stated that specific adaptation is connected with large genotypic PCA1 scores to environments with PCA1 scores of the same sign. For example, G14, which had a positive IPCA1 score of 0.49, was specifically adapted to E4 with a positive IPCA1 score of 0.92. This is also true with negative IPCA1 scores and many genotypes in this study demonstrated the same relationship. Environments E6, E4 and E3 had the highest effect on GEI. Genotypes that were clustered on the centre of origin such as G15 and G2 had their mean performance closer to the grand mean hence revealing general adaptation to the testing environments. Genotype G27 was specifically adapted to E6 and G23 similarly was adapted to E2. This is so because they had acute angles to the test environments in context. Fox *et al*. (1997) explained that the smaller the angle between the genotypes and respective environments the more the genotypes are specifically adapted to that particular environment.

Environments E1, E6, E5 and E3 were the most informative since their GEI variation was larger as depicted by the length of their vectors. GEI indicates differences in adaptation and it can be exploited by selecting for specific adaption if the study is repeated over years (Yan *et al.*, 2007) hence in this study broad adapted genotypes G4, G2 and G11 can be recommended for cultivar release and minimize specific adaptation selection since the study was not repeated over years.

The “which won where” pattern helps in visualising the possible existence of mega environments in multi environmental trials and shows the best performing genotype in each environment (Kaya *et al.*, 2006). Either the genotypes on the vertex of the polygon formed were the best or poorest in the sectors and designated environments they fell in (Yan *et al.*, 2007). Genotype G11 won in environment E5, for environments E3, E2, E6, E4 and E1 the winners were G22, G17, G4, G6 and G4, respectively. For G3 and G10 vertex lines had no environment in their sector implying that they performed poorly across all locations as also illustrated by a study conducted by Asfaw *et al.* (2009). G15 and the others close to the point of origin of the biplot had their mean performance close to the grand mean, which means their performance across the locations had the same response.

An ideal genotype is defined as having the highest mean performance and is stable even though such type of genotypes may not exist but in ranking and evaluation of genotypes, they can be used as a reference. Genotype G4 was almost closer to the ideal genotype hence it is the best performer in terms of high yield and stability. This concept is also applicable to environments; the environment in a concentric ring closer to the ideal environment is ideal among the test environments. In this case, environment E6 had the most representative and discriminating ability since it had an acute angle to the average environment axes (AEA) and it is suitable for selection of generally adapted genotypes. Environments E3 and E1 had larger angles between their vectors and the AEA hence these environments were suitable for selection of specifically adapted genotypes. From the test environments, no genotype was the best in all environments indicating a crossover GEI. This situation could also have been influenced by the climatic conditions of the environments since they had different temperature, rainfall and soil conditions.

Yield performance and stability was defined using the first and second interaction principal component axis scores of all test locations symbolised by a small circle. Two lines pass through the origin of the biplot, the first one is the average environment axis, and this has an arrow pointing to greater GEI effect and reduced stability. The second one, the ordinate of the AEC runs perpendicular to the AEC (Kaya *et al*., 2006). Genotypes G4, G11 to G17 had their performances above average hence can be recommended for all test locations provided that there is optimal climatic conditions and improved management practices. The AEC ordinate separates genotypes with below average yields to those with above average yields. For selection, genotypes with above average means could be selected, that is, from G4 to G17 and discard the others such as G10 which had low yields (non-adaptable) and unstable. Stability should be considered during selection even among the genotypes with performances above average. For example, G4, G2 and G12 were high yielding and more stable since their vectors where closer to the AEC while G17, G22 and G11 were high yielding but more variable across the test environments.

E2 had the lowest IPCA value of -30.31, which symbolised low interaction with the climatic condition, and E5 had the highest IPCA score. G14 had the highest IPCA score of 0.49 while C3 had the lowest value of -0.85 indicating that it was more stable across locations. This is in contrast with GGE analysis, which indicates G2 was the most stable because it had the shortest vector from the AEC.

2.5 Conclusion

The present study revealed that soybean yield was significantly affected by genotype, environment and genotype by environment as revealed through the AMMI analysis. GGE analysis using the “which won where” pattern identified genotypes with specific adaptation such as G11 to environment E5. It also identified genotypes G4, G2 and G11 as having general adaptation and high yielding. For the ideal genotype, G4 was identified since it had the longest vector hence it was high yielding and highly stable hence it can be recommended for cultivar release in the tested environments. E1 and E6 were both in the same sector hence one environment could be used for selection of the other environment to reduce the cost of the breeding trials. G22 and G11 had low yield mean performance but stable hence they can be further improved by using them as parents in another breeding pipeline by crossing to high yielding genotypes.

REFERENCES

Abate, T., A.D. Alene, D. Bergvinson, B. Shiferaw, S. Silim, A. Orr. 2012. Tropical grain legumes in Africa and south Asia: knowledge and opportunities. International Crops Research Institute for the Semi-Arid Tropics.

Alghamdi, S.S. 2004. Yield stability of some soybean genotypes across diverse environments. Pakistan Journal of Biological Sciences 7: 2109-2114.

Allard, R.W., & Bradshaw, A. (1964). Implications of Genotype-Environmental Interactions in Applied Plant breeding 1. Crop Science 4(5): 503-508.

Asfaw, A., F. Alemayehu, F. Gurum and M. Atnaf. 2009. AMMI and SREG GGE biplot analysis for matching varieties onto soybean production environments in Ethiopia. Scientific Research and Essays 4: 1322-1330.

Atnaf, M., S. Kidane, S. Abadi and Z. Fisha. 2013. GGE biplots to analyze soybean multi-environment yield trial data in north Western Ethiopia. Journal of Plant Breeding and Crop Science 5: 245-254.

Baker, J., L. Allen, K. Boote, P. Jones and J. Jones. 1989. Response of soybean to air temperature and carbon dioxide concentration. Crop Science 9: 98-105.

Bhartiya, A., J. Aditya, K. Singh, J. Purwar and A. Agarwal. 2017. AMMI & GGE biplot analysis of multi environment yield trial of soybean in North Western Himalayan state Uttarakhand of India. Legume Research: An International Journal 40(2): 56-65

Branquinho, R.G., J.B. Duarte, P.I.M.d. Souza, S.P.d. Silva Neto and R.M. Pacheco. 2014. Environmental stratification and optimization of a multi-environment trial net for soybean genotypes in Cerrado. Pesquisa Agropecuária Brasileira 49: 783-795.

Ceccarelli, S. 1989. Wide adaptation: how wide? Euphytica 40: 197-205.

Chipeta, M.M., R. Melis, P. Shanahan, J. Sibiya and I.R. Benesi. 2017. Genotype x environment interaction and stability analysis of cassava genotypes at different harvest times, Journal of Animal and Plant Sciences 27: 901-919.

Cooper, M. and G.L. Hammer. 1996. Plant adaptation and crop improvement. IRRI.

Eberhart, S.t. and W. Russell. 1966. Stability parameters for comparing varieties. Crop Science 6: 36-40.

Fox, J. (1997). Applied regression analysus, linear models, and related methods. Sage Pulications, Inc.

Gauch, H. and R. Zobel. 1989. Accuracy and selection success in yield trial analyses. Theoretical and Applied Genetics 77: 473-481.

Gauch, H.G. 2006. Statistical analysis of yield trials by AMMI and GGE. Crop Science 46: 1488-1500.

Kaya, Y., M. Akçura and S. Taner. 2006. GGE-biplot analysis of multi-environment yield trials in bread wheat. Turkish Journal of Agriculture and Forestry 30: 325-337.

Mohammadi, R., M. Armion, E. Zadhasan, M.M. Ahmadi and A. Amri. 2017. The use of AMMI model for interpreting genotype × environment interaction in durum wheat. Experimental Agriculture: 1-14. doi:10.1017/S0014479717000308.

Mohammadi, R., R. Haghparast, A. Amri and S. Ceccarelli. 2010. Yield stability of rainfed durum wheat and GGE biplot analysis of multi-environment trials. Crop and Pasture Science 61: 92-101.

Murithi, H., F. Beed, M. Soko, J. Haudenshield and G. Hartman. 2015. First report of Phakopsora pachyrhizi causing rust on soybean in Malawi. Plant Disease 99: 420-420.

Nzuve, F., S. Githiri, D. Mukunya and J. Gethi. 2013. Analysis of genotype x environment interaction for grain yield in maize hybrids. Journal of Agricultural Science 5: 75-85.

Oliveira, V.M., O.T. Hamawaki, A.O. Nogueira, L.B. Sousa, F.M. Santos and R.L. Hamawaki. 2016. Selection for wide adaptability and high phenotypic stability of Brazilian soybean genotypes. Genetics and Molecular Research 15: 13. doi:10.4238/gmr.15017843.

Payne, R.W., Harding, S.A., Murray, D.A., Soutar, D.M., Baird, D.B., Glaser, A.L., Channing, I.C., Welham, S.J., Gilmour, A.R., Thompson, R. & Webster, R. 2010. The Guide to Genstat Release 14 Part 3: Statistics.Hemel Hempstead UK: VSN International.

Romagosa, I., P. Fox, L.G. Del Moral, J. Ramos, B.G. del Moral, F.R. De Togores. 1993. Integration of statistical and physiological analyses of adaptation of near-isogenic barley lines. Theoretical and Applied Genetics 86: 822-826.

Sibiya, J., P. Tongoona, J. Derera and N. van Rij. 2012. Genetic analysis and genotype× environment (G× E) for grey leaf spot disease resistance in elite African maize (*Zea mays* L.) germplasm. Euphytica 185: 349-362.

Yan, W. and M.S. Kang. 2002. GGE biplot analysis: A graphical tool for breeders, geneticists, and agronomists. CRC press.

Yan, W., M.S. Kang, B. Ma, S. Woods and P.L. Cornelius. 2007. GGE biplot vs. AMMI analysis of genotype-by-environment data. Crop Science 47: 643-653.

CHAPTER 3

CORRELATION, PATH COEFFICIENT AND GENOTYPE BY TRAIT ASSOCIATION ANALYSIS AMONG ELITE SOYBEAN LINES ACROSS ENVIRONMENTS

ABSTRACT

Grain yield in general is a complex trait with low heritability and is dependent upon different variables hence indirect selection through other component traits would be an essential strategy to improve the efficiency of selection. It is important to measure the contribution of each trait to grain yield through partitioning them into direct and indirect effects and graphically show interrelationships among the traits. The study was conducted on 30 genotypes to determine the correlation, and path analyses of grain yield and graphically visualise trait relationships. Significant differences at P<0.001among genotypes were observed for all traits studied. Correlation coefficient between grain yield with early vigour and plant height illustrated a strong positive relationship. Days to 50% flowering, days to 50% podding days to maturity hundred seed weight and had a negative correlation coefficient with grain yield. Path analysis indicated that plant height had a positive direct effect on yield while early vigour and days to 50% flowering had negative indirect effects on yield. The genotype by trait (GT) biplot graphically showed consistent trait relationships and identified G4 (TGX 2001-3FM), G27 (TGX 2001-26DM) and G9 (TGX 2002-3DM) as genotypes that can be used as parents in breeding programmes for yield improvement.

Key words: Soybean, trait profiles, correlation, path coefficient analysis

3.1 Introduction

Soybean is an economically important leguminous crop that is grown for its oil and protein products (Tefera *et al.*, 2009). It also has medicinal properties and is a source of raw material for industries. Grain yield is a quantitative trait, which is dependent on a number of other characters. For yield improvement in breeding programmes, the study of direct and indirect effects of yield and its attributing components can be used as a baseline for selection for grain yield through the closely associated characters (Malik *et al.*, 2007).

In plant breeding programmes, there is a common goal to identify traits that positively contribute to high yield, therefore, it is critical to study traits in a crop and identify those that contribute to the trait of interest (Kinfe *et al.*, 2015). Various statistical methodologies have been employed to understand genotype and trait interactions, which leads to having the same or similar output (Akcura, 2011).

Genotype plus genotype by environment (GGE) biplots have been used for analysing multi environment trials. This concept can be extended to analyse data for multiple traits across locations. The genotypes are used in the place of environments and the multiple traits are used as testers. With this data, the genotype by trait (GT) biplot is constructed, which presents a visual display showing association of traits and genotypes. This is an effective tool that graphically summarizes the genotype by trait data, visualises relationships among the measured traits and visualise the performance of genotypes based on the traits, which influence selection of potential parents (Yan and Tinker, 2005). In addition, it helps to identify less important traits that do not contribute directly to the trait of interest. Genotype by trait association has been used in soybean yield analysis by Yan and Kang (2002) who reported that one genotype performed the best across all locations.

Agrama (1996) stated that the proficiency of any breeding programme relies on how large the association is between yield and its components. Most of the main breeding traits have negative correlations existing among them hence selection for a single trait is very difficult. Correlation coefficients are useful in quantifying the trait associations in terms of size and magnitude. However, they might be misleading if there is a large correlation that is a result of indirect effect of a trait. In soybean, scientists have used path analysis to partition correlations into direct and indirect traits (Malik *et al*., 2007). The objectives of this study were to determine genotype by trait associations and multiple trait relationships across six locations.

3.2. Materials and methods

3.2.1. Genotypes and evaluation environments

Twenty-six elite soybean lines along with four checks were used to generate data used in this analysis during the 2016/2017 growing season at six locations spread over three countries (Malawi, Mozambique and Zambia). The list of genotypes used and the geographical information of each location (environment) are shown in Tables 3.1 and 3.2, respectively. Of the 30 genotypes evaluated, TGX 2002-3DM was early maturing and determinate in growth habit while the rest were medium maturing and indeterminate in growth habit.

Table 3.1 Genotypes evaluated and their agronomic characteristics

Genotype	Genotype code	Source	Maturity	Growth habit
TGx 2001-12FM	G1	IITA	M	I
TGx 2001-16FM	G2	IITA	M	I
TGx 2001-24DM	G3	IITA	M	I
TGx 2001-3FM	G4	IITA	M	I
TGx 2001-4FM	G5	IITA	M	I
TGx 2001-5FM	G6	IITA	M	I
TGx 2002-12FM	G7	IITA	M	I
TGx 2002-1FM	G8	IITA	M	I
TGx 2002-3DM	G9	IITA	E	D
TGx 2002-3FM	G10	IITA	M	I
TGx 2002-4FM	G11	IITA	M	I
TGx 2002-9FM	G12	IITA	M	I
TGx 2001-19FM	G13	IITA	M	I
TGx 2001-7FM	G14	IITA	M	I
TGx 2002-10FM	G15	IITA	M	I
TGx 2001-14FM	G16	IITA	M	I
TGx 2001-27DM	G17	IITA	M	I
TGx 2001-6DM	G18	IITA	M	I
TGx 2002-6DM	G19	IITA	M	I
TGx 2001-13FM	G20	IITA	M	I
TGx 2001-14DM	G21	IITA	M	I
TGx 2001-15DM	G22	IITA	M	I
TGx 2001-21DM	G23	IITA	M	I
TGx 2001-21FM	G24	IITA	M	I
TGx 2002-11FM	G25	IITA	M	I
TGx 2001-1FM	G26	IITA	M	I
TGx 2001-26DM	G27	IITA	M	I
Tikolore	G28	IITA	M	I
NASOKO	G29	IITA	M	I
SC SERENADI	G30	IITA	M	I

IITA = International institute for tropical agriculture, M = medium, E = early, I = intermediate, D = Determinant

Table 3.2 Trial locations and their geographical information

Country	Location and Environment Code	Altitude	Latitude	Longitude	Min Temp	Max Temp
		m.asl	^{0}S	^{0}E	^{0}C	^{0}C
Malawi	Chitedze E6	1149	13.9815	33.6372	18	29
	Bvumbwe E2	1146	15.917	35.067	14	25
Zambia	IITA SARAH E4	1199	15.302	27.574	15	30
	Kabwe E5	1182	14.4285	28.4514	17	28
Mozambique	Nampula E1	360	15.1266	39.2687	19	34
	Gurue E3	788	15.4914	37.0125	19	29

m. asl = metres above sea level

3.2.2. Design of trials and agronomic management

The soybean genotypes were planted in a 6 x 5 alpha lattice design with three replications. The plots comprised of four rows of 4 m in length, 0.5 m between rows and 0.05 m intra-row spacing. Data were collected from the two middle rows, which were considered as a net plot leaving one row and 0.5 m on either sides as borders. Manual and chemical weeding was done to mitigate weeds and inorganic fertilizers were applied at all locations depending on the results of soil analysis.

3.2.3. Data collection and analysis

Data on plant characteristics listed in Table 3.3 were collected for all genotypes at all evaluation sites.

Table 3.3 Agronomic traits and their description

Trait	Acronym	Description
Early Vigor	EV	Visual assessment at seedling stage, Scoring 1-good, 3-Intermediate, 5-poor
Days to flowering	DFFL	Count number days after sowing to 50% flowering of plants in a plot
Days to podding	DPD	Count days after sowing to 50% podding of plants in a plot
Days to maturity	DM	Count days after sowing to full maturity
Hundred seed weight	HSW	100 seeds were counted and weighed
Grain yield	GY	Dry and weight of total grains per plot
Plant height	PLHT	The average height of 5 plants from soil level to shoot tip at maturity

Pearson correlation and path coefficient analysis

The Pearson correlation coefficient was computed using the formula in SAS 9.4 software as indicated in Equation 3.5.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \text{ ....Equation 3.1}$$

Where; r is the Pearson coefficient correlation, x is the dependent variable, y is the independent variable while n is the sample size

For path coefficient analysis, PROC CALIS was implemented using SAS 9.4. This analysis revealed direct and indirect effects on the primary trait that was grain yield. The model used as suggested by Akintunde (2012) and is indicated in Equation 3.2

$$y = a + b_1X_1 + b_2X_2 + b_3X_3 + U \text{ ....Equation 3.6}$$

Where: y is the dependable variable (GY) while $a + b_1X_1 + b_2X_2 + b_3X_3 + U$ are the correlation variables with the assumption that each variable is independently contributing to the dependent variable y

Analysis of Variance

The data were subjected to analysis of variance (ANOVA) across locations.

The model (equation 3.3) for the combined ANOVA of multi-environment trials was used in GenStat 17th edition (Payne et al., 2014). The model includes additive terms for main effects of genotype and environment, as well as the genotype by environment interaction term.

$$\mathrm{Y}_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \mathrm{E}_{ij} \text{ ....Equation 3.3}$$

Where: Yij is the yield of the genotype i in environment j and kth replication; μ is overall yield mean, αi and βj are genotypic and environmental effect, (αβ)ij is the effect of interaction between the ith genotype and jth environment, ∈ij is the mean random error of the ith genotype and ej environment.

Genotype by trait analysis

The genotype by trait model used was adopted from Yan and Rajcan (2002) and is presented in Equation 3.4 as follows:

$$\frac{\mathrm{T}_{ij}-\beta_{ij}}{s_j}=\sum_{n=1}^{2}\lambda_n\xi_{in}\eta_{jn}+\varepsilon_{ij}=\sum_{n=1}^{2}\xi^{*}_{in}\eta^{*}_{jn}+\varepsilon_{ij}$$ ....Equation 3.7

Where:

T_{ij} = The mean value of genotype i for trait j

β_j = The average value of all genotypes for trait j

S_j = The standard deviation of trait j among genotype means

λ_n = The singular value for Principal Component (PCn)

ξ_{in} = The PCn score for genotype i

η_{jn} = The PCn score for trait j

ε_{ij} = The residual associated with genotype i in trait j

To achieve symmetric scaling between the genotype scores and the trait scores the singular value λ_n has to be absorbed by the singular vector for genotypes ξ_{in} and that for traits η_{jn}. That is, $\xi^{*}_{in} = \lambda^{0.5}{}_{n}\, \xi_{in}$ and $\eta^{*}_{jn} = \lambda^{0.5}{}_{n}\, \eta_{jn.}$ Only PC1 and PC2, are retained in the model because such a model tends to be the best for extracting pattern and rejecting noise from the data. The GT biplot is generated by plotting ξ^{*}_{i1} and ξ^{*}_{i2} against η^{*}_{j1} and η^{*}_{j2}, respectively, so that each genotype or trait is represented by a marker in the biplot. In the GT biplot, a vector is drawn from the biplot origin to each marker of the traits to facilitate visualization of the relationships between and among the traits.

3.3. Results

3.3.1. Combined analysis of variance and agronomic performance of genotypes

The results of the combined ANOVA for each agronomic trait were highly significant at P<0.001 (Table 3. 3) for genotype, site and site by genotype interaction.

Table 3.4 Analysis of variance for seven traits across six locations

Source	DF	EV	DFFL	DM	PHLT	SWT	GY	DPD
Site	5	3.58***	22.56***	117.32***	428.83***	56.1***	1181557.39***	118.72***
Rep(Site)	12	0.4 ns	2.63ns	4.54ns	36.59ns	6.21ns	81474.29ns	12.52**
Block(Site*Rep)	72	0.4ns	4.07**	4.67ns	23.82ns	5.21***	87538.79ns	9.04**
Genotype	29	2.57***	82.31***	263.93***	678.26***	7.69***	675579.76***	90.22***
Site*Genotype	145	1.77***	54.54***	212.17***	377.53***	10.06***	668209.26***	82.79***
Error	276	0.37	2.38	3.49	21.34	2.6	93149.8	5.71

***Significant at P<0.001, **Significant at P<0.01, ns = not significant, DF = Degrees of freedom, EV = early vigour, DFFL = days to 50% flowering, DM = days to maturity, PHLT= plant height, SWT= 100 seed weight, GY= grain yield, DPD= Days to 50% podding

3.3.2. Mean comparisons for the evaluated genotypes

None of the genotypes was best for all the traits. Table 3.4 presents a comparison of genotype mean for each trait across the locations using Tukey test at 5% probability level. Plant height ranged from 45.185 cm to 69.56 cm with genotype G10 having the lowest value and G27 having the highest value. Days to maturity ranged from 88.8 days to 104.8 days with G5 recording the longest time to maturity and G9 recording the shortest time to maturity. For hundred seed weight, the range was 16.23 g to 19.02 g with G3 as the highest and C3 as the lowest. The highest number of days to podding recorded was 56.57 days and the lowest being 45.56; G4 had the lowest number of days and G23 was the highest. G2 had the lowest score of 2.1 for early vigour and G17 had the highest score of 3.81. Grain yield ranged from 1309.7 kg/ha to 2080.5 kg/ha with G4 as the best and G24 as the lowest yielding genotype.

Table 3.5. Mean comparisons of the evaluated genotypes based on seven agronomic traits

Genotype	EV	DFFL	DM	PHLT	HSW	GY	DPD
G1	2.58cdefghi	39.07bcd	100.80bcd	47.13hi	16.60bc	1501.1defghi	52.44bcdef
G2	2.1i	37.32cdefgh	95.19ghi	52.489efgh	16.91abc	1991.2ab	50.8bcdefgh
G3	2.27fghi	36.85efghi	97.39fg	53.42defgh	19.02a	1605.7bcdefghi	51.04bcdefg
G4	2.43efghi	34.01klmn	91.13klm	63.98ab	17.24abc	2080.5a	45.56l
G5	2.9bcdefghi	37.16cdefgh	104.88a	50.56efghi	17.61abc	1833.8abcdef	53.78abc
G6	2.98abcdefgh	36.63efghi	98.84cdef	60.03bc	17.18abc	1748.9abcdefgh	49.15fghijk
G7	3.36abc	38.06cdefg	98.53def	49.82efghi	17.75abc	1447.6efghi	51.41bcdefg
G8	3.25abcdef	37.07defgh	97.23fgh	49.72efghi	16.73bc	1942.7abc	51.49bcdefg
G9	3.26abcde	33.11mn	88.86m	65.41ab	17.563abc	1775.7abcdefgh	46.37kl
G10	2.97bcdefgh	36.06ghij	99.74bcde	45.19i	17.35abc	1369.5hi	50.15defghij
G11	3.3abcd	36.18fghij	93.69ijk	47.46hi	16.99abc	1774.5abcdefgh	49.84efghij
G12	2.52cdefghi	38.17cdefg	94.74hij	47.69fghi	17.21abc	1524.3cdefghi	52.33bcdef
G13	2.73bcdefghi	33.44lmn	93.63ijkl	53.76cdefg	16.98abc	2041.4a	47.67hijkl
G14	2.92bcdefghi	41.43a	107.23a	51.43efghi	18.35abc	1606.8bcdefghi	53.47abcd
G15	3.08abcdefg	41.09ab	92.98ijkl	50.35efghi	18.45abc	1472.3efghi	52.35bcdef
G16	3.43ab	39.30abc	99.17cdef	63.14ab	17.94abc	1906.6abcd	53.81ab
G17	3.81a	38.95cd	101.34bc	59.55bcd	18.26abc	1735.7abcdefghi	52.58bcde
G18	3.07abcdefg	37.13defgh	94.86ghij	54.00cdef	16.89abc	1734.6abcdefghi	50.47cdefghi
G19	2.41fghi	32.58n	93.75ij	68.85a	17.99abc	1660abcdefghi	47.49ijkl
G20	2.52cdefghi	38.20cdefg	98.76cdef	48.31fghi	18.73ab	1830.3abcdef	50.05efghij
G21	2.42efghi	35.30hijkl	92.46jkl	47.46ghi	17.37abc	1872.3abcde	47.6hijkl
G22	2.48defghi	34.20jklmn	92.42jkl	55.64cde	17.63abc	2030ab	49.97efghij
G23	2.73bcdefghi	41.09ab	98.74def	59.71bcd	18.54ab	1398.8ghi	56.57a
G24	3.01abcdefgh	38.22cdef	100.51bcde	49.96efghi	17.29abc	1309.7i	51.52bcdefg
G25	2.69bcdefghi	38.57cde	102.19b	51.29efghi	17.79abc	1550.7cdefghi	52.77bcde
G26	2.2fhi	33.35lmn	92.37jkl	54.91cde	16.88abc	1835.5abcdef	48.97ghijk
G27	2.4ghi	34.83ijklm	91.02lm	69.57a	16.62bc	1811.2abcdefg	46.98jkl
C1	2.53cdefghi	37.73cdefg	98.14ef	52.06efgh	17.42abc	1653.4abcdefghi	52.1bcdefg
C2	2.64bcdefghi	37.00defgh	95.40ghi	53.84cdefg	18.69ab	1489.7defghi	52.23bcdefg
C3	2.61bcdefghi	37.21cdefgh	93.68ijk	48.27fghi	16.232c	1439.2fghi	52.04bcdefg
Mean	2.78	36.9	96	54.16	17.5	1699	50
Range	2.1 - 3.8	32.5 - 41.4	88.8 - 107.2	45.1 - 69.6	16.2 - 19.1	1309.6 - 2080.4	45.5 - 56.5
SE	0.6	1.5	1.8	4.6	1.6	305.2	2.3
P-Value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
CV%	21.8	4	1.9	8.5	9.1	17.9	4.7

EV = Early vigour, DFFL = days to flowering, DM = days to maturity, PHLT = plant height, HSW, = hundred seed weight, GY = grain yield, DPD = days to podding. Means followed by the same letter in columns indicate that they do not differ at 5% probability by Tukey test

3.3.3. Correlation analysis

Plant height, early vigour and days to 50% flowering were highly significant and correlated with grain yield. Plant height had a weak, positive and significant correlation (r =0.27) with yield. Hundred seed weight, days to maturity and days to podding had negative non-significant correlations with yield (Table 3.6).

Table 3.6 Pearson's correlation coefficient between the agronomic traits

	EV	DFFL	DPD	DM	PLHT	HSW	GY
EV	1	-0.32***	-0.21***	-0.17***	0.12**	0.08	0.13**
DFFL		1	0.73***	0.59***	-0.28***	-0.07	-0.30***
DPD			1	0.57***	-0.29***	-0.11*	-0.31
DM				1	-0.2***	-0.001	-0.23
PLHT					1	0.13**	0.27***
HSW						1	-0.02
GY							1

EV = Early vigour, DFFL = days to flowering, DM = days to maturity, PHLT = plant height, HSW, = hundred seed weight, GY = grain yield, DPD = days to podding.

3.3.4. Path coefficient analysis

From the path analysis results for grain yield (Table 3.7), the coefficients were partitioned into direct and indirect effects through different characters affecting the grain yield component. The direct effects of plant height and early vigour were positive while for days to flowering, days to podding, days to maturity and hundred seed weight had negative effects. The highest direct effect was observed on plant height followed by early vigour. The lowest direct effect was through days to podding. It was also noted that high indirect effects were observed in early vigour while the other traits including days to flowering and days to podding had negative indirect effects on yield via days to maturity.

Table 3.7 Path analysis illustrating direct and indirect effects on grain yield

	Effects on GY		
	Total	Direct	Indirect
EV	0.1284**	0.0353	0.0932***
DFFL	-0.2847***	-0.0979	-0.1868***
DPD	-0.2013**	-0.1614**	-0.0399
DM	-0.0367	-0.0344	-0.00232**
PHLT	0.191***	0.1985***	-0.00745
HSW	-0.0685	-0.0685	0.0000
	Effects on DM		
EV	-0.18***	0.000617	0.1794***
DFFL	0.5871***	0.3533***	0.2338***
DPD	0.3155***	0.3155***	0.0000

3.3.5. Genotype by trait analysis

The genotype by trait biplot has an ability to visually show comparisons of genotypes for multiple traits and their associations.

3.3.5.1. Identification of the best genotypes based on multiple traits

The GT biplot was constructed using data from 7 agronomic traits of 30 genotypes across six environments. Eighty percent of the total variation was explained by the biplot from the standardised mean data with PC1 contributing 66.12% and PC2 accounting for 14.48%. The polygon view was divided into eight sectors with the traits falling in four of the sectors. The traits days to maturity, days to flowering, days to podding fell in the same sector, hundred seed weight and early vigour fell in the same sector while PHLT and GY each fell in its own sector. G14, G9, G4, G26 and G21 were on the vertex of the polygon view in the GT biplot (Figure 3.1). The genotypes on the vertices performed either the best or poorly in one or more of the traits. G4 was the highest performing genotype for grain yield while G9 was the highest recorded for plant height. G17 and G23 were found to be highly associated with hundred seed weight and early vigour. G14 was observed to be highly associated with days to flowering, days to maturity and days to podding. Genotypes G1, G30 and G26 exhibited low associations with all the traits used in the study.

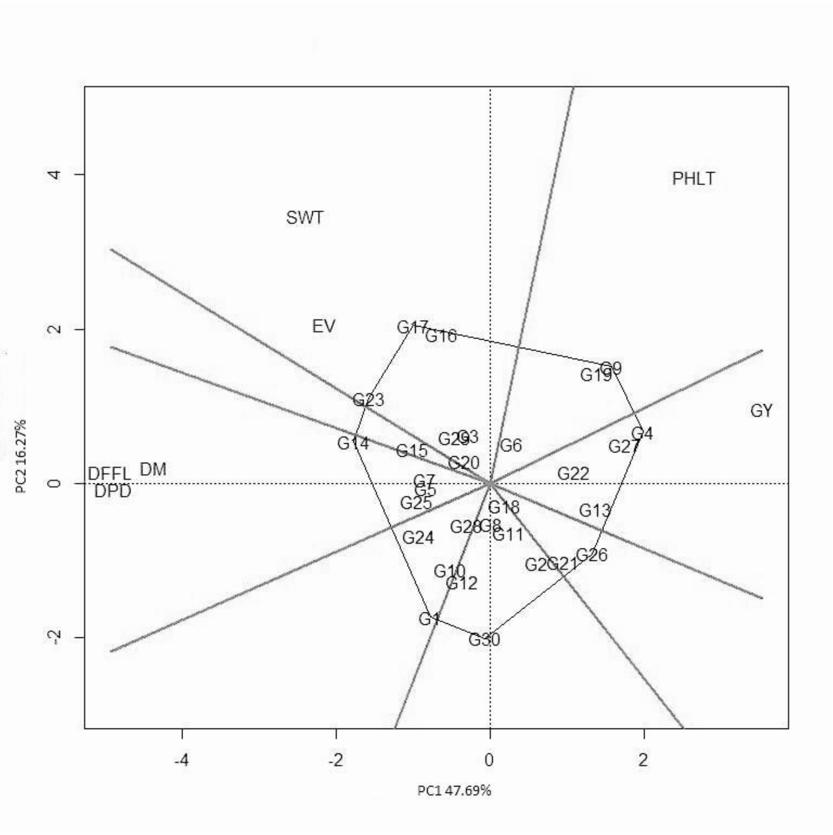


Figure 3.1 "Which is best for what" plot. PHLT = plant height GY =grain yield. DFFL = days to flowering, DPD= days to podding, EV= early vigour, DM= days to maturity, SWT= hundred seed weight

3.3.5.2. Relationships among traits

Figure 3.2 illustrates trait relationships. PHLT and DM had the longest vectors compared to the EV and SWT, which had the shortest vectors. Traits GY and PHLT, DPD and DFFL, EV and DM had acute angles between their vectors. The angles between these vectors indicate the correlation between the test traits. Angles between vectors that are less than 90^0 shows that there is a high correlation between the traits as observed among: grain yield and plant height, days to podding and days to flowering; days to maturity, early vigour and days to maturity, hundred seed weight and early vigour. Vectors for early vigour and plant height were at right angles indicating that they were independent of each other. The angle between grain yield and days to maturity and between days to podding and days to flowering was more than 90^0 illustrating a negative correlation among the traits. Both the angle and length of the trait vectors indicate the similarity between the test traits.

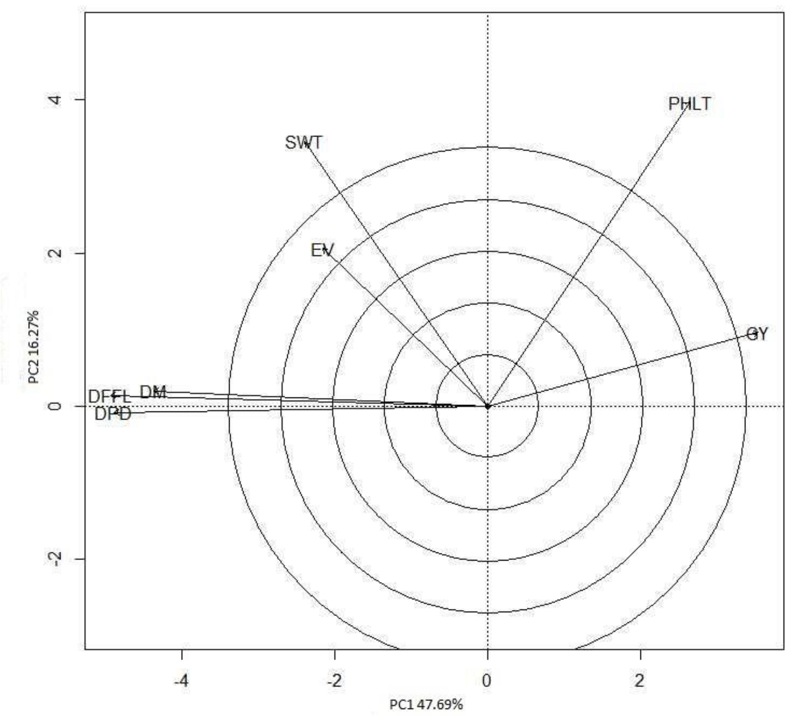


Figure 3.2 Trait relationships among genotypes

3.3.5.3. Comparison of trait profiles of two specific genotypes

Trait profiles of two genotypes can be easily compared on the GT biplot. Two genotypes, G4 (the highest yielding genotype) and G24 (the lowest yielding genotype) are compared in Figure 3.3. G4, the highest yielding genotype, had a higher association with traits such as grain yield, plant height while for genotype G24, the lowest yielding, had a close association with days to podding and early vigour.

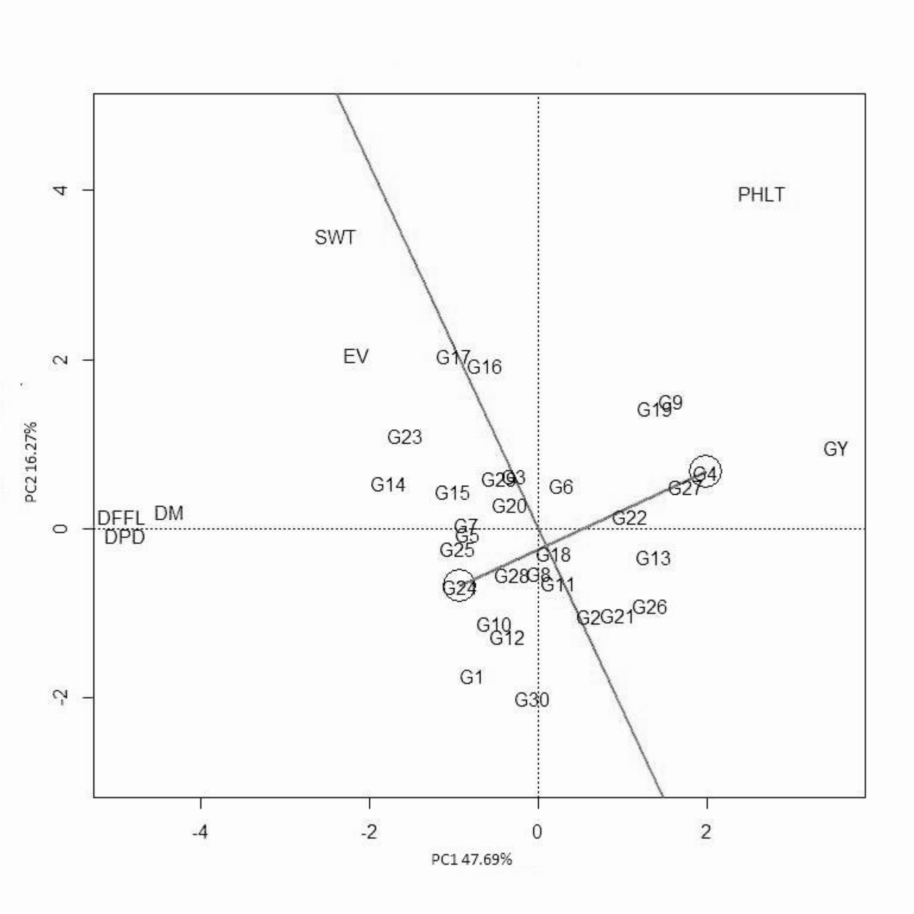


Figure 3.3 Genotype comparisons based on trait profiles

3.4. Discussion

From the evaluation of the genotypes across the environments, the genotype by trait biplots illustrated visual comparisons of genotypes and relationships among the traits. The biplot explained 80% of the total variation. These high percentages of variation indicate accuracy among the measured entities (Badu-Apraku and Akinwale, 2011). G4 had the highest value for grain yield and G19 for plant height hence these genotypes can be used as parents if these traits are to be improved in other cultivars such as G21 and G1 that were observed to be poor. G4, G9, G27 and G19 had the highest values for both grain yield and plant height hence these genotypes can also be used for yield improvement and plant height can be used for indirect selection for yield. Across the diverse environments, genotypes (G4, G19, G19 and G22) which were associated with high grain yield and inherent traits could be used as parents in cultivar development programmes targeting stable and high yielding genotypes.

Plant height and days to maturity had the largest effect on yield compared to the other traits these traits had the longest vector indicating that they have an ability to discriminate among the genotypes. Days to maturity had a negative effect on yield because of the obtuse angle between days to maturity and grain yield vectors while plant height had a positive correlation as illustrated by the acute angle between plant height and grain yield. Plant height and days to maturity were independent of each other because the angle between their vectors was 90^0. Agrama (1996) stated that the proficiency of any breeding programme relies on how large the association is between yield and its components. Most of the main breeding traits have negative correlations existing among them hence selection for a single trait is very difficult.

The high correlations between plant height and grain yield implies that either grain yield or plant height can be used as a selection tool for the trait, this is in agreement with the Pearson's correlation which also shows that plant height and grain yield had a highly significant positive correlation. This emphasizes the importance of GT identifying redundant traits. It reduces the cost of measuring the traits in experiments without compromising precision (Odewale *et al.*, 2013). The distance measured from the centre of origin of the biplot to a genotype marker, indicates the difference of the genotype from the averaged genotype of all traits that is hypothetically positioned at the centre of origin (Yan and Frégeau-Reid, 2008). Genotypes G14, G19, G21 and G6 had the longest vectors implying that they have large values for a particular trait. These may be best genotypes or not but can be used as parents for some traits.

The path coefficient study revealed that plant height and early vigour had high and positive direct effects on grain yield and days to maturity had negative direct effect on grain yield. This result is consistent with Arshad *et al.* (2006). Days to podding had high negative indirect effects on yield via days to maturity. From the analysis, it is recommended that for yield improvement selection can be based on plant height and number of days to maturity. This is also consistent with the results obtained from the GT biplot analysis. However, GT is more informative and interpretable due to the graphical presentation that enhances patterns among the traits. Days to 50% podding, days to 50% flowering and 100-seed weight had negative direct effects on yield via days to maturity. This implies that selection of yield based on these traits might lead to the loss of soybean yield.

A study conducted at the National Agriculture Research center in Islamabad in summer by Anwarmalik et al. (2007) on 27 genotypes of soybean to define the correlation and path analysis of yield and its associated components, revealed that days to flowering, days to podding, and plant height had a maximum direct contribution to yield. However, this was the opposite of the findings in the study except for one trait; plant height, which also showed a direct contribution to grain yield. Hence, the traits with direct contribution in this study; plant height and early vigour can be utilised as a selection criteria in improving the soybean genotypes.

3.5. Conclusion

The GT biplot explained a high percentage of the total variation in the data set. It also described interactions among traits that gave similar results with Pearson’s correlations. PHLT had a positive direct effect to yield hence it can be used as a selection trait that contributes to yield that is, the more the values for plant height the higher the yield. It is more logical that the longer a plant takes to mature the more the yield, the opposite of this was observed in this study. G4, G9, and G19 can be recommended to be used as parents in other breeding programmes for yield improvement.

REFERENCES

Agrama, H. 1996. Sequential path analysis of grain yield and its components in maize. Plant Breeding 115: 343-346.

Akçura, M. 2011. The relationships of some traits in Turkish winter bread wheat landraces. Turkish Journal of Agriculture and Forestry 35: 115-125.

Akintunde, A. 2012. Path analysis step by step using excel. Journal of Technical Science and Technologies 1: 09-15.

Anwarmalik, M. F., Ashraf, M., Qureshi, A. S. and Ghafoor, A. (2007). Assessment of Genetic Variability, Correlation and Path Analyses for Yield and its Components in Soybean, Pakistan Journal of Botany 39 (2): 405-413.

Arshad, M., N. Ali and A. Ghafoor. 2006. Character correlation and path coefficient in soybean *Glycine max* (L.) Merrill. Pakistan Journal of Botany 38: 121-130.

Badu-Apraku, B. and R.O. Akinwale. 2011. Cultivar evaluation and trait analysis of tropical early maturing maize under Striga-infested and Striga-free environments. Field Crops Research 121: 186-194. doi:https://doi.org/10.1016/j.fcr.2010.12.011.

Ebdon, J. and H. Gauch. 2002. Additive main effect and multiplicative interaction analysis of national turfgrass performance trials. Crop Science 42: 497-506.

Kinfe, H., G. Alemayehu, L. Wolde and Y. Tsehaye. 2015. Correlation and Path Coefficient Analysis of Grain Yield and Yield Related Traits in Maize (*Zea mays* L.) Hybrids, at Bako, Ethiopia. Journal of Biology, Agriculture and Healthcare 5: 15-22.

Malik, M.F.A., M. Ashraf, A.S. Qureshi and A. Ghafoor. 2007. Assessment of genetic variability, correlation and path analyses for yield and its components in soybean. Pakistan Journal of Botany 39: 405-413

Odewale, J., C. Agho, C. Ataga, E. Okolo, C. Ikuenobe and M. Ahanon. 2013. Genotype by trait relations of yield and Other Physiological Traits of coconut (*Cocos nucifera* L.) hybrids based on GT Biplot. Merit Research Journal of Agricultural Science and Soil Sciences 1: 042-049.

Tefera, H., A. Kamara, B. Asafo-Adjei and K. Dashiell. 2009. Improvement in grain and fodder yields of early-maturing promiscuous soybean varieties in the Guinea Savanna of Nigeria. Crop Science 49: 2037-2042.

Yan, W. and J. Frégeau-Reid. 2008. Breeding line selection based on multiple traits. Crop Science 48: 417-423.

Yan, W. and M.S. Kang. 2002. GGE biplot analysis: A graphical tool for breeders, geneticists, and agronomists. CRC press.

Yan, W. and I. Rajcan. 2002. Biplot analysis of test sites and trait relations of soybean in Ontario. Crop Science 42: 11-20.

Yan, W. and N.A. Tinker. 2005. An integrated biplot analysis system for displaying, interpreting, and exploring genotype× environment interaction. Crop Science 45: 1004-101

CHAPTER 4

ASSESSMENT OF GENETIC DIVERSITY IN TROPICAL SOYBEAN LINES USING SINGLE NUCLEOTIDE POLYMORPHISMS

ABSTRACT

Genetic diversity is an important element in plant breeding and the basis for genetic improvement. Knowledge of how diverse the genotypes are genetically, is useful for the conservation, utilization and management of germplasm collections. This study was conducted to estimate the genetic diversity among 48 soybean lines from the International Institute for Tropical Agriculture (IITA). The lines were evaluated for genetic diversity using 348 Kompetitive Allele Specific Polymerase Chain Reaction (KASP) genotyping assays based on competitive allele-specific PCR which enables bi-allelic scoring of single nucleotide polymorphism markers. The obtained bi-allelic data was analysed for diversity using GenAlex software version 6.5. The average gene diversity ranged from 0.42 to 0.55 with an average of 0.47. The genetic distance ranged from 0.61 to 0.87. The polymorphic information content ranged from 0.44 to 0.50 with a mean of 0.48. Genotypes TGX 2002-3DM and TGX 2002-3FM had the highest genetic distance between them indicating that they were highly diverse. The analysis of molecular variance indicated highly significant differences at F=0.001 with among individuals, among populations and within individuals contributing 45%, 28% and 26%, respectively, to the total variance. The 48 soybean lines were clustered in three main groups. The study indicated that genetic diversity exists among the IITA tested lines. The information obtained from the study can be fully utilised in future soybean breeding programmes through crossing of diverse parents in order to introgress new alleles to develop improved cultivars.

Key words: Single nucleotide polymorphisms, genetic diversity, soybean

4.1. Introduction

Soybean (*Glycine max* L) is a self-pollinating crop that belongs to the Leguminosae family. It is believed that China is the center of origin and diversity. It was cultivated in China as early as the 11th century and probably domestication was done earlier before that time. Sailors from the 17th century brought soybean to Europe then later it was introduced to Africa. Soybean, among the oil crops has gained popularity for oil and protein products and it is only second to groundnuts.

Genetic diversity is an important element in plant breeding for genetic improvement. This concept is defined as biodiversity found within plant species that form the basis of plant breeding (Dong *et al.*, 2001). Most plant breeding programmes aim at selecting genotypes that are superior within a diverse population hence the knowledge of genetic diversity among the soybean genotypes would play a vital role in selection. The genetically diverse genotypes can be selected as parents and are likely to give high heterotic effects and the segregating population would have high frequency of desirable genotypes to select from (Hipparagi *et al.*, 2017).

Various methods have been implored to assess genetic variation including evaluating morphological traits (Oliveira and Valls, 2003), pedigree analysis (Sneller, 1994), biochemical analysis (Javaid *et al.*, 2004), and DNA markers have been used recently (Feng *et al.*, 2008). Morphological traits are somehow able to differentiate genotypes using conventional breeding methods. This, however, brings difficulties when selecting among cultivars that have the same parents since they are closely related, their morphological traits would probably be similar. Therefore, studying genetic diversity using molecular markers is the most effective way to distinguish these genotypes (Rodrigues *et al.*, 2017).

Several molecular markers have been used in soybean genetic diversity studies and they have proved to be effective in distinguishing cultivars (Doldi *et al.*, 1997). Molecular markers are highly polymorphic and reproducible hence, they are the best technique to use in plants with narrow genetic variation. In addition, the advantage of molecular markers compared to morphological assessment, is that, they determine the extent of genetic relatedness among cultivars. Tan *et al.* (2012) stated that molecular markers offer an opportunity for direct selection at DNA level since these genetic markers are based on individual nucleotide sequence variation. Random Amplified Polymorphic DNA markers (RAPD) have been used in genetic diversity studies but their repeatability is low hence limiting their application. Simple sequence repeats (SSR) markers have also been widely used due to their large number of alleles per locus. However, SSR genotyping is expensive and time consuming when dealing with large populations. Molecular markers used to study genetic diversity in soybean include SSRs (Priolli *et al.*, 2002), RAPDs (Brown-Guedira *et al.*, 2000), amplified fragment length polymorphisms (AFLPs) (Rocha *et al.*, 2015), restriction fragment length polymorphisms RFLPs (Diers *et al.*, 1992), and single nucleotide polymorphisms (SNPs) (Zhu *et al.*, 2003).

SNP markers have proved to be the most abundant sources of DNA polymorphisms. They are defined as the single DNA base differences between DNA fragments including insertions and deletion (Zhu *et al.*, 2003). With these properties, they can be easily used for genetic and association mapping, genetic diversity studies and genome wide selection. SNP genotyping has been in several other studies such as cereals, cowpea and pea (Tan *et al.*, 2012). However, there is not much information on SNP markers in soybean and therefore, the study was conducted to assess the genetic diversity of tropical soybean lines sourced from IITA using SNP markers.

4.2. Materials and methods

4.2.1. Experimental material

Forty-six elite soybean lines along with two checks were used to generate data used in this analysis during the 2016/2017 growing season. The list of genotypes used and their agronomic characteristics are shown in Table 4.1

Table 4.1 Genotype evaluated and their characteristics

Genotype	Genotype code	Source	Maturity	Growth Habit
TGX 2001-26DM	G1	IITA	M	I
TGX 2001-3FM	G2	IITA	M	I
TGX 2001-19FM	G3	IITA	M	I
TGX 2001-15DM	G4	IITA	M	I
TGX 2002-1FM	G5	IITA	M	I
TGX 2001-6DM	G6	IITA	M	I
TGX 2001-4FM	G7	IITA	M	I
NASOKO	G8	IITA	M	I
TGX 2001-21FM	G9	IITA	M	I
TGX 2001-5FM	G10	IITA	M	I
TGX 2002-3DM	G11	IITA	M	I
SC SERENADI	G12	IITA	M	I
TIKOLORE	G13	IITA	M	I
TGX 2001-7FM	G14	IITA	M	I
TGX 2002-12FM	G15	IITA	M	I
TGX 2002-11FM	G16	IITA	M	I
TGX 2001-13FM	G17	IITA	M	I
TGX 2002-3FM	G18	IITA	M	I
TGX 2002-4FM	G19	IITA	M	I
TGX 2001-24DM	G20	IITA	M	I
TGX 2001-12FM	G21	IITA	M	I
TGX 2002-10FM	G22	IITA	M	I
TGX 2001-27DM	G23	IITA	M	I
TGX 2001-14DM	G24	IITA	M	I
TGX 2001-14FM	G25	IITA	M	I
TGX 2001-1FM	G26	IITA	M	I
TGX 2001-16FM	G27	IITA	M	I
TGX 2001-21DM	G28	IITA	M	I
TGX 2002-9FM	G29	IITA	M	I
TGX 2002-6DM	G30	IITA	E	D
TGX 2014-24FM	G31	IITA	E	D
TGX 2014-32FM	G32	IITA	E	D
TGX 2001-11DM	G33	IITA	E	D
TGX 2014-33FM	G34	IITA	E	D
TGX 2002-14DM	G35	IITA	E	D
TGX 1991-22F	G36	IITA	E	D
TGX 2014-5GM	G37	IITA	E	D
TGX 2014-42FM	G38	IITA	E	D
TGX 2014-28FM	G39	IITA	E	D
TGX 2014-4FM	G40	IITA	E	D
TGX 2001-16DM	G41	IITA	E	D
TGX 2014-31FM	G42	IITA	E	D
TGX 2002-22DM	G43	IITA	E	D

Genotype	Genotype code	Source	Maturity	Growth Habit
TGTX 1989-60F	G44	IITA	E	D
TGX 2001-18DM	G45	IITA	E	D
TGX 2001-10DM	G46	IITA	E	D
TGX 2001-26FM	G47	IITA	E	D
TGX 2014-15FM	G48	IITA	E	D

M= medium, E= early, I= Indeterminate, D= Determinate

4.2.2. Greenhouse nursery

The 48 soybean lines were planted in a greenhouse at Chitedze (IITA-Malawi) in polythene tubes. Three seeds per genotype were planted per tube and replicated two times. Peat was used as the growth media.

4.2.3. DNA sampling and isolation

At four weeks after planting, eight leaf discs were harvested from three plants per polythene tube. These were used for the DNA extraction. Sampling kit was obtained from LGC Genomics Laboratory, United Kingdom and it included a 96-well plate, cutting mat and leaf-cutting tool. Leaf samples from the same genotype were placed in a specific well position of the plate, each strip was sealed using perforated trip cap. The desiccant sachet was placed directly on top of the strip cap-sealed tubes and the plastic lid was replaced on top. The storage rack was secured by using an elastic band and was placed inside a sealable plastic bag. The sealed bag was placed into the plant kit box and the samples were shipped to LGC Genomics Laboratory for genotyping in the United Kingdom.

4.2.4. SNP selection and amplification

In compliance with the protocol supplied by LGC Genomics Laboratory, Kompetitive Allele Specific Polymerase Chain Reaction (KASP) genotyping assays were used. These were based on competitive allele-specific PCR and enable bi-allelic scoring of single nucleotide polymorphisms (SNPs) and insertion and deletions (Indels) at specific loci.

The SNP-specific KASP Assay mix and the universal KASP Master Mix (supplied at 2X concentration) were used. KASP Master Mix contains Taq polymerase enzyme and passive reference dye, 5-carboxy-X-rhodamine, succinimidyl ester (ROX) and $MgCl_2$ in an optimized buffer solution. The two mix were added to DNA samples then a thermal cycling was performed, followed by an end-point fluorescent read. Allele-specific primers each harbouring a unique tall sequence that correspond with a universal fluorescence resonant energy transfer (FRET) cassette; one labelled with FAMTM dye and the other with HEXTM dye were used. During thermal cycling, the relevant allele-specific primer would bind to the template and elongate, thus attaching the tail sequence to the newly synthesized strand. The complement of the allele-specific tail sequence was then generated during subsequent rounds of PCR, enabling the FRET cassette to bind to the DNA. Bi-allelic discrimination was achieved through the competitive binding of the two allele-specific forward primers. If the genotype was heterozygous, a mixed fluorescent was generated. If the genotype at a given SNP was homozygous, only one of the two possible fluorescent signal was generated (Figure 4.1). For the current study, 348 SNP markers were used.

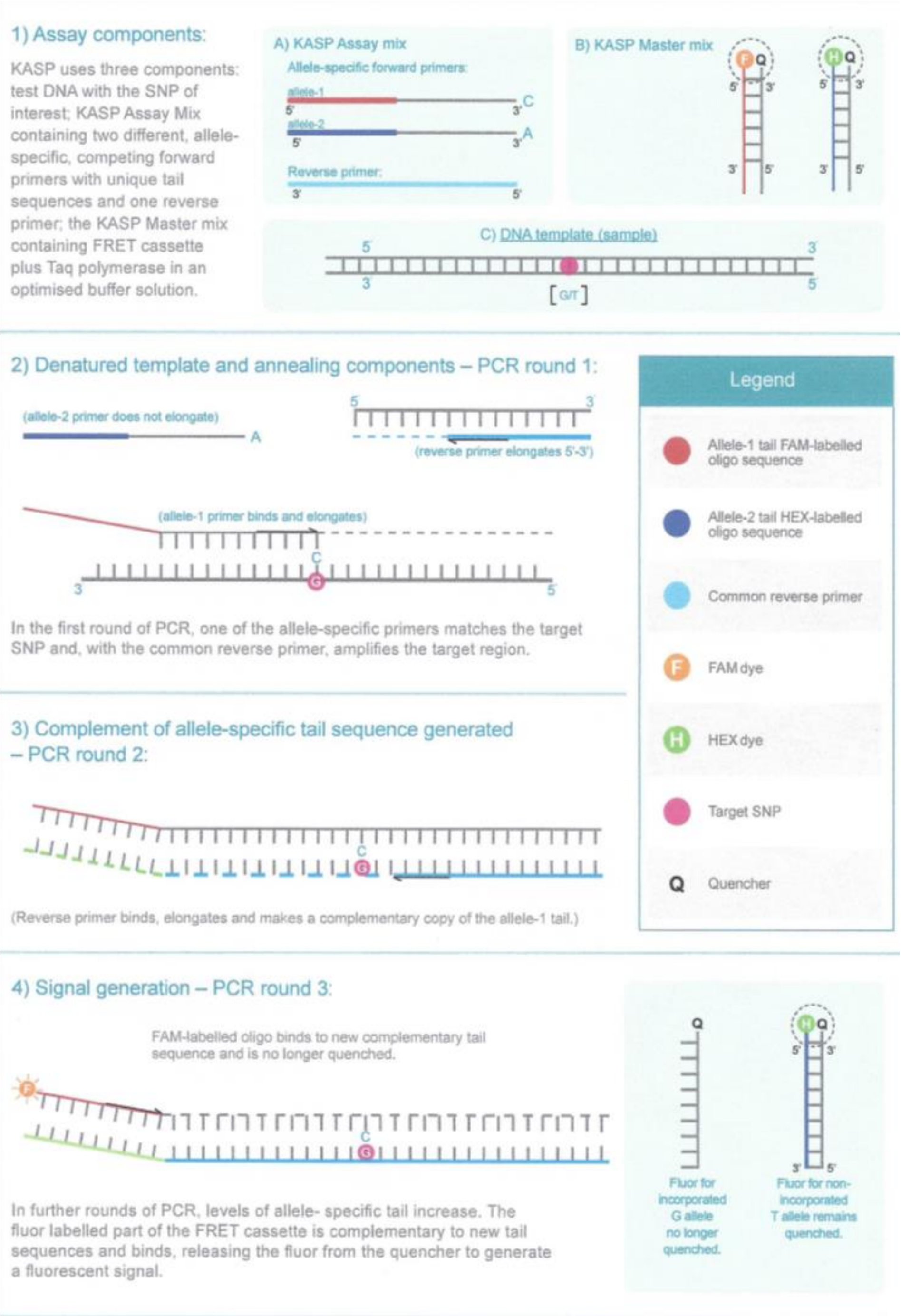

1) Assay components:
KASP uses three components: test DNA with the SNP of interest; KASP Assay Mix containing two different, allele-specific, competing forward primers with unique tail sequences and one reverse primer; the KASP Master mix containing FRET cassette plus Taq polymerase in an optimised buffer solution.
A) KASP Assay mix
Allele-specific forward primers:
allele-1
5'
3'
C
allele-2
5'
3'
A
Reverse primer:
3'
5'
B) KASP Master mix
F
Q
H
Q
5'
3'
3'
5'
C) DNA template (sample)
5'
3'
3'
5'
[G/T]
2) Denatured template and annealing components – PCR round 1:
(allele-2 primer does not elongate)
A
5
3
(reverse primer elongates 5'-3')
(allele-1 primer binds and elongates)
C
G
3
5
In the first round of PCR, one of the allele-specific primers matches the target SNP and, with the common reverse primer, amplifies the target region.
Legend
Allele-1 tail FAM-labelled oligo sequence
Allele-2 tail HEX-labelled oligo sequence
Common reverse primer
F FAM dye
H HEX dye
Target SNP
Q Quencher
3) Complement of allele-specific tail sequence generated – PCR round 2:
C
G
(Reverse primer binds, elongates and makes a complementary copy of the allele-1 tail.)
4) Signal generation – PCR round 3:
FAM-labelled oligo binds to new complementary tail sequence and is no longer quenched.
F
C
G
In further rounds of PCR, levels of allele- specific tail increase. The fluor labelled part of the FRET cassette is complementary to new tail sequences and binds, releasing the fluor from the quencher to generate a fluorescent signal.
Q
H
Q
5'
3'
3'
5'
Fluor for incorporated G allele no longer quenched.
Fluor for non-incorporated T allele remains quenched.

Figure 4.1 KASP procedure (Protocol supplied by LGC Genomics Laboratory, UK)

4.2.5. Data Analysis

The genotypic data were analysed using GenAlex software version 6.5 to obtain information of genetic diversity within and among populations. Genetic diversity parameters, such as number of alleles per locus (N_a), number of effective alleles per locus (N_e), observed (H_o) and expected (H_e) heterozygosity, and Shannon's Information Index (I) were calculated using GenAlex version 6.5 Peakall and Smouse (2006) according to the protocol described by Nei and Li (1979). The number of polymorphic loci was estimated for each predetermined group based on the inferred population using Structure (Pritchard *et al.*, 2000). Further, an indirect estimate of the level of gene flow (N_m) was calculated using the formula: N_m = 0.25 (1 – FST/FST) using GenAlex. The F-statistics such as genetic differentiation (F_{ST}), fixation index or inbreeding coefficient (F_{IS}), and overall fixation index (F_{IT}) were calculated according to Wright's original derivation (Wright, 1978). Polymorphic information content (PIC) was calculated using the formula: PIC =1 - ΣP_{ij}^2, where P_{ij} is the frequency of jth allele of the ith locus. Nei’s unbiased genetic distance was also estimated to determine the degree of population differentiation among the study material. Nei’s unbiased genetic distance and identity were estimated according to Nei and Li (1979) using GenAlex.

The genotypic data were used to obtain a dissimilarity matrix using the Jaccard index. The matrix was used to run a cluster analysis. Cluster analysis was done based on neighbour-joining algorithm using the un-weighted pair group method using arithmetic average (UPGMA) in DARwin 5.0 software (Perrier and Jacquemoud-Collet, 2006). A dendrogram was then generated on the dissimilarity matrix. To investigate the genetic relationships among accessions, genetic distances between all pairs of individual accessions were estimated to draw a dendrogram. Bootstrap analysis was performed for node construction using 10,000 bootstrap values.

The Bayesian genotypic clustering approach of STRUCTURE 2.3.4 (Pritchard *et al.,* 2000) was used to determine the population structure existed with the genotypes. An admixture model with independent allele frequencies, without prior population information, was used to simulate the population. Each individual was grouped in a given cluster using ‘membership coefficient’ for each cluster interpreted as a probability of membership. To assign individual genotypes to a given population and for optimal alignment of genotypes, 10 replicates structure analysis were conducted. The computer programme CLUMPP (Jakobsson and Rosenberg, 2007) was used to determine the genotype membership. The structure analysis result was visualized by the online genetic software STRUCTURE HARVESTER (Earl, 2012).

The number of genotypes that represented the populations were unbalanced; allelic richness was corrected for sample size differences and estimated by using the rarefaction method implemented in HP-Rare 1.0 (Kalinowski, 2005).

4.3. Results

From the data analysis, it showed that from the 348 SNP markers, 199 markers were observed to be polymorphic. The PIC ranged from 0.41 to 0.5 with an average of 0.48 (Table 4.2). Approximately more than 85% of the markers used had high PIC values of more than 0.42, which implies that they have a high ability to distinguish the variation in the genotypes using the SNP markers. The 48 genotypes used in the study were divided in to four populations as shown in Table 4.3. Population 3 had the highest number of genotypes of 17 followed by population 1 with 15 genotypes and populations 2 and 4 comprised of 9 and 7 genotypes respectively.

Table 4.2 Genetic diversity within and among 48 soybean genotypes based on 348 SNPs markers

Chromosome	SNP marker used	Polymorphic SNPs	**N_e**	**H_o**	**H_e**	**F_{IS}**	**PIC**
1	6	3	1.88	0.15	0.47	0.69	0.466
2	10	8	1.96	0.19	0.49	0.61	0.489
3	31	19	1.91	0.12	0.47	0.75	0.475
4	20	10	1.96	0.21	0.49	0.58	0.485
5	17	12	1.84	0.14	0.45	0.69	0.447
6	39	21	2.06	0.13	0.51	0.75	0.500
7	19	10	2.04	0.15	0.50	0.68	0.498
8	16	5	1.82	0.20	0.45	0.54	0.443
9	12	10	2.10	0.21	0.53	0.60	0.500
10	17	15	2.02	0.09	0.50	0.81	0.499
11	14	5	2.08	0.13	0.52	0.75	0.500
12	31	17	2.01	0.14	0.50	0.71	0.498
13	7	6	2.12	0.11	0.53	0.78	0.500
14	20	8	1.99	0.18	0.50	0.64	0.492
15	13	7	1.96	0.15	0.49	0.68	0.484
16	10	4	1.85	0.06	0.42	0.89	0.412
17	3	3	2.21	0.13	0.55	0.77	0.500
18	27	19	2.01	0.14	0.50	0.71	0.494
19	12	6	1.82	0.10	0.45	0.77	0.445
20	24	11	1.90	0.13	0.48	0.72	0.470
Overall mean	348	199	1.68	0.12	0.37	0.65	0.480
SE			0.02	0.01	0.01	0.02	0.005

N_e= number of effective alleles per locus; H_o= observed gene diversity within genotypes; H_e= average gene diversity within genotypes; F_{IS}= inbreeding coefficient; PIC= polymorphic information content; SE= standard deviation

Table 4.3 Population membership inferred by Structure

Population 1	Population 2	Population 3	Population 4
TGx 2001-6DM	TGx 2002-1FM	TGx 2001-26DM	TGx 2001-4FM
TGx 2002-12FM	Tikolore	TGx 2001-3FM	NASOKO
TGx 2002-11FM	TGx 2001-13FM	TGx 2001-19FM	TGX 1991-22F
TGx 2002-3FM	TGx 2001-24DM	TGx 2001-15DM	TGX 2001-16DM
TGx 2002-4FM	TGx 2001-14DM	TGx 2001-21FM	TGX 2014-31FM
TGx 2001-12FM	TGx 2001-16FM	TGx 2001-5FM	TGTX 1989-60F
TGx 2001-14FM	TGx 2001-21DM	TGx 2002-3DM	TGX 2001-18DM
TGX 2014-24FM	TGx 2002-6DM	SC SERENADI	
TGX 2014-32FM	TGX 2014-42FM	TGx 2001-7FM	
TGX 2001-11DM		TGx 2002-10FM	
TGX 2014-33FM		TGx 2001-27DM	
TGX 2014-5GM		TGx 2001-1FM	
TGX 2014-28FM		TGx 2002-9FM	
TGX 2002-22DM		TGX 2002-14DM	
TGX 2001-10DM		TGX 2014-4FM	
		TGX 2001-26FM	
		TGX 2014-15FM	

From the diversity data, the values ranged from 0.06 to 0.21 with an average of 0.12. Approximately 60% of the markers had values greater than 0.12. The highest number of the effective alleles per locus was 1.85 in population 4 and the lowest value observed was 1.56 in population 1 with an average of 1.69 (Table 4.4). The highest inbreeding coefficient of 0.69 was revealed in population 4 while the lowest of 0.33 was observed in population 2 with a mean of 0.56. The highest average gene diversity within genotypes per population was 0.43 from population 4 while the lowest observed value of 0.33 from population 1.

Table 4.4 Genetic diversity within and among 48 soybean genotypes classified into four population based on structure analysis

Pop	N	N_a	N_e	I	H_o	H_e	F_{IS}	P_a
Population 1	14	2.30	1.56	0.52	0.09	0.33	0.67	12
Population 2	9	2.26	1.67	0.59	0.24	0.40	0.33	9
Population 3	17	2.25	1.68	0.57	0.15	0.37	0.56	13
Population 4	8	2.33	1.85	0.65	0.10	0.43	0.69	24
Overall mean	12	2.29	1.69	0.58	0.15	0.38	0.56	-
SE	0.13	0.02	0.02	0.01	0.01	0.01	0.01	-

N= number of genotypes within population; N_a= number of alleles per locus N_e= number of effective alleles per locus; H_o= observed gene diversity within genotypes; H_e= average gene diversity within genotypes; F_{IS}= inbreeding coefficient; P_a= number of private alleles; SE= standard deviation

Table 4.5 showed that there was a significant genetic distance between the genotypes among the populations. The highest genetic distance was observed between populations 1 and 3 with a value of 0.48. The lowest recorded value was 0.14 that was between populations 2 and 3. The generated dendrogram (Figure 4.2) shows the genetic relationships among the genotypes. It also illustrated that the markers used in the study were effective in distinguishing the genotypes into three different clusters as indicated using three different colours: black, red and blue in the figure.

The analysis of molecular variance showed that, among populations contributed 28% of the total variation. The variation among individuals of the total population was 45% indicating high differentiation among them and the variation accounted within individuals was 23% (Table 4.6).

Table 4.5 Pair-wise estimates of genetic differentiation (FST) (above diagonal off brackets), gene flow (Nm) (above diagonal within brackets); genetic distance GD (lower diagonal off brackets) and genetic identity (GI) (lower diagonal within brackets)

	Population 1	Population 2	Population 3	Population 4
Population 1		0.19 (1.07)	0.26 (0.70)	0.15 (1.43)
Population 2	0.30 (0.74)		0.11 (2.14)	0.18 (1.17)
Population 3	0.48 (0.61)	0.14 (0.87)		0.16 (1.30)
Population 4	0.23 (0.80)	0.30 (0.74)	0.25 (0.78)	

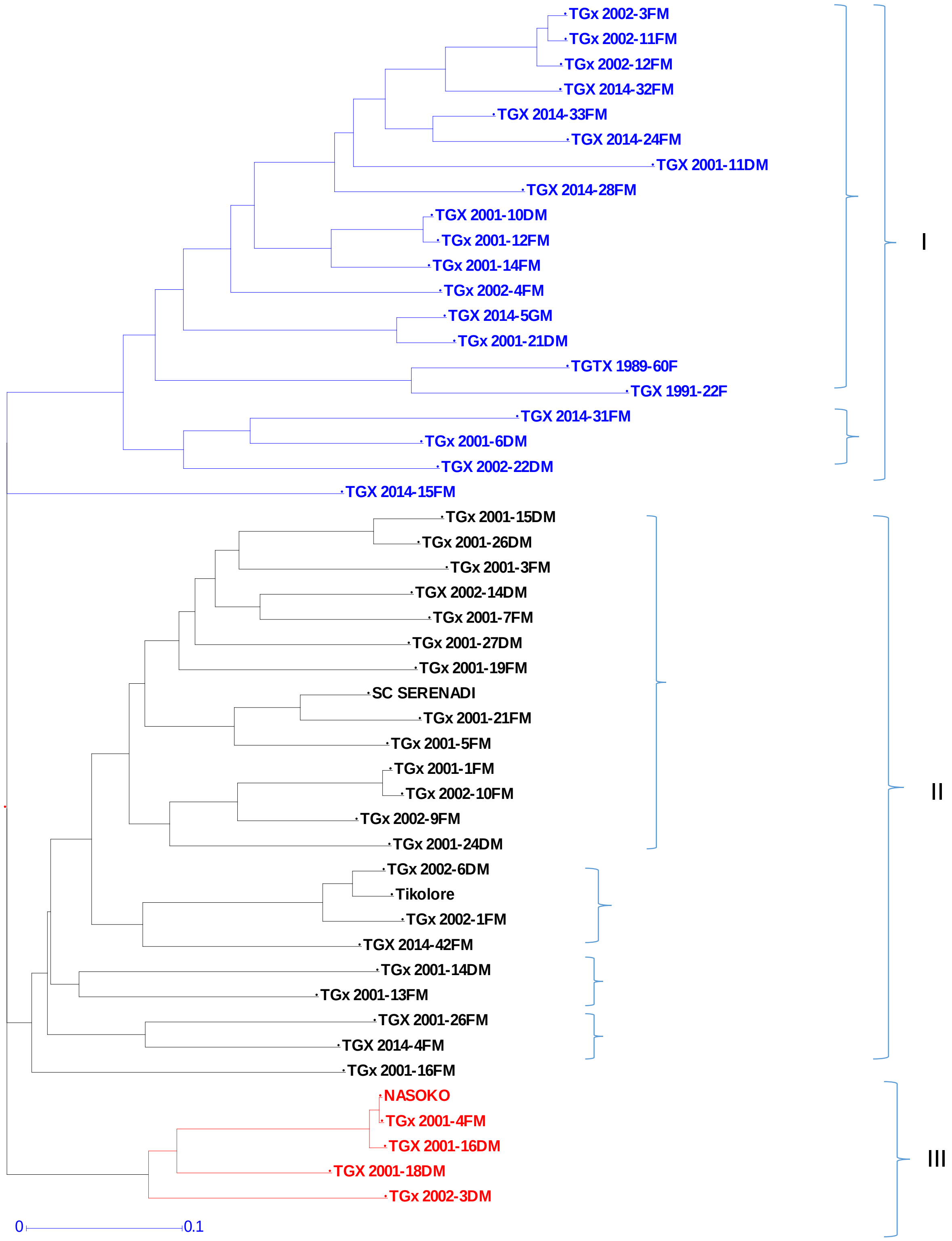


Figure 4.2 Dendrogram showing genetic relationship among 48 soybean genotypes tested using 348 SNP markers. The different colours indicate the clustering patterns among the genotypes

Table 4.6 Analysis of molecular variance (AMOVA) among 48 soybean genotypes classified based on Structure analysis

Source	df	SS	MS	Est. Var.	Per. Var.	F-Statistics
Among populations	3	1234.57	411.52	15.02	28%	F_{ST} = 0.001
Among individuals	44	2739.87	62.27	24.10	45%	F_{IS} = 0.001
Within individuals	48	675.50	14.07	14.07	26%	F_{IT} = 0.001
Total	95	4649.94		53.19	100%	

DF= degree of freedom, SS= sum of squares, MS= mean sum of squares, Est. var. = estimated variance, Per. Var. = percentage variation

4.4. Discussion

Using the 348 SNP markers, the 48 soybean accessions where clustered into three main genetic groups. This entails that the markers were effective in discriminating the soybean lines. This clustering based on genetic similarity from this study would help in selection of genetically diverse genotypes to be used as parents for superior recombinants in soybean breeding programmes.

The average gene diversity ranged from 0.42 to 0.55 and the polymorphic information content ranged from 0.44 to 0.5. From the previous genetic diversity studies, according to Mulato *et al.* (2010) a high PIC value of 0.92 indicates that there is great diversity between the accessions. Thus, the primers used were highly informative compared to the ones observed in this study. This might be due to the use of rare SNPs that are not present in our accessions. Botstein *et al.* (1980) indicated a scale of mean PIC value >0.5 is highly informative, 0.25-0.50 reasonably informative and <0.25 is slightly informative, hence the set of SNPs used in this study were reasonably informative and reliable.

A difference between the total number of alleles and the number of effective alleles for all the chromosomes among populations was observed. The total number of alleles was higher than the number of effective alleles. This was observed because the frequency of the alleles at a single locus were distributed differently among genotypes. The highest genetic distance was observed between population 2 and 3 with a value of 0.87 indicating that genotypes from these populations belong to different genetic clusters. The gene flow observed in population 3 and 1 with a value of 0.70 was the lowest compared to the other estimates. This value is less than 1 hence it has the potential to significantly reduce the loss of genetic diversity by preventing the effect of genetic drift (Aguilar *et al.*, 2008). Genotypes observed with high genetic distance can be used as potential parents in developing superior cultivars as they are expected to have greater genetic variations (Li *et al.*, 2008). Generally, among soybean genotypes exists low genetic diversity, which is said to be because of gene flow as a result of seed exchange between farmers.

The analysis of molecular variance (AMOVA) indicated highly significant differences from all sources of variation, and this implies that there were significant genetic differences among the genotypes subjected to this analysis. The variance among populations was significant and contributed 28% of the total variation, while the variance among individuals was significantly high contributing 45% of the total variation. The variance within individuals was significant and contributed 26% of the total variation.

4.5. Conclusion

Results of the genetic diversity analysis indicate that 199 polymorphic SNPs out of the 348 SNP markers that were used genetically distinguished the genotypes. The set of SNP markers used have the ability to accurately estimate genetic diversity among soybean genotypes used. This study revealed wide genetic diversity among soybean genotypes; therefore, a breeding programme can be initiated between genotypes from different clusters to exploit available genetic diversity.

REFERENCES

Aguilar, R., M. Quesada, L. Ashworth, Y. Herrerias-Diego and J. Lobo. 2008. Genetic consequences of habitat fragmentation in plant populations: susceptible signals in plant traits and methodological approaches. Molecular Ecology 17: 5177-5188.

Botstein, D., White, R.L., Skolnick, M., & Davis, R.W., 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. American Journal of Human Genetics 32(3): 314.

Brown-Guedira, G., J. Thompson, R. Nelson and M. Warburton. 2000. Evaluation of genetic diversity of soybean introductions and North American ancestors using RAPD and SSR markers. Crop Science 40: 815-823.

Diers, B., P. Keim, W. Fehr and R. Shoemaker. 1992. RFLP analysis of soybean seed protein and oil content. Theoretical and Applied Genetics 83: 608-612.

Doldi, M.L., J. Vollmann and T. Lelley. 1997. Genetic diversity in soybean as determined by RAPD and microsatellite analysis. Plant Breeding 116: 331-335.

Dong, Y.S., B.C. Zhuang, L.M. Zhao, H. Sun and M.Y. He. 2001. The genetic diversity of annual wild soybeans grown in China. Theoretical and Applied Genetics 103: 98-103. doi:10.1007/s001220000522.

Earl, D.A. 2012. STRUCTURE HARVESTER: a website and programme for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources 4: 359-361.

Feng, C., A. Hou, P. Chen, B. Cornelious, A. Shi and B. Zhang. 2008. Genetic diversity among popular historical Southern US soybean cultivars using AFLP markers. Journal of Crop Improvement 22: 31-46.

Hipparagi, Y., R. Singh, D.R. Choudhury and V. Gupta. 2017. Genetic diversity and population structure analysis of Kala bhat (*Glycine max* (L.) Merrill) genotypes using SSR markers. Hereditas 154: 9. doi:10.1186/s41065-017-0030-8.

Jakobsson, M. and N.A. Rosenberg. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23: 1801-1806.

Javaid, A., A. Ghafoor and R. Anwar. 2004. Seed storage protein electrophoresis in groundnut for evaluating genetic diversity. Pakistan Journal of Botany 36: 25-30.

Kalinowski, S.T. 2005. HP-rare 1.0: a computer program for performing rarefaction on measures of allelic richness. Molecular Ecology Resources 5: 187-189.

Li, Y., R. Guan, Z. Liu, Y. Ma, L. Wang, L. Li. 2008. Genetic structure and diversity of cultivated soybean (*Glycine max* (L.) Merr.) landraces in China. Theoretical and Applied Genetics 117: 857-871. doi:10.1007/s00122-008-0825-0.

Mulato, B.M., M. Möller, M.I. Zucchi, V. Quecini and J.B. Pinheiro. 2010. Genetic diversity in soybean germplasm identified by SSR and EST-SSR markers. Pesquisa Agropecuária Brasileira 45: 276-283.

Nei, M. and W. H. Li. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proceedings of the National Academy of Sciences 76: 5269-5273.

Oliveira, M.A.P. and J.F.M. Valls. 2003. Morphological characterization and reproductive aspects in genetic variability studies of forage peanut. Scientia Agricola 60: 299-304.

Peakall, R. and P.E. Smouse. 2006. GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecology Resources 6: 288-295.

Perrier, X. and J. Jacquemoud-Collet. 2006. DARwin software: Dissimilarity analysis and representation for windows. Website http://darwin. cirad. fr/darwin [accessed 1 March 2013].

Priolli, R.H.G., C.T. Mendes-Junior, N.E. Arantes and E.P.B. Contel. 2002. Characterization of Brazilian soybean cultivars using microsatellite markers. Genetics and Molecular Biology 25: 185-193.

Pritchard, J.K., M. Stephens and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155: 945-959.

Rocha, C.M.L., G.R. Vellicce, M.G. García, E.M. Pardo, J. Racedo, M.F. Perera, 2015. Use of AFLP markers to estimate molecular diversity of *Phakopsora pachyrhizi*. Electronic Journal of Biotechnology 18:439-444. doi:http://dx.doi.org/10.1016/j.ejbt.2015.06.007.

Rodrigues, J.I.d.S., K.M.A. Arruda, C.D. Cruz, E.G.d. Barros, N.D. Piovesan and M.A. Moreira. 2017. Genetic divergence of soybean genotypes in relation to grain components. Ciência Rural 47-58

Sneller, C.H. 1994. Pedigree analysis of elite soybean lines. Crop Science 34: 1515-1522.

Tan, H., M. Tie, Q. Luo, Y. Zhu, J. Lai and H. Li. 2012. A review of molecular makers applied in Cowpea (*Vigna unguiculata* L. Walp.) Breeding. Journal of Life Sciences 6: 1190.

Wright, S. 1978. Variability within and among populations. Vol 4. Evolution and the Genetics of Populations. Genetics Selection Evolution 10: 75-81.

Zhu, Y., Q. Song, D. Hyten, C. Van Tassell, L. Matukumalli, D. Grimm. 2003. Single-nucleotide polymorphisms in soybean. Genetics 163: 1123-1134.

CHAPTER 5

OVERVIEW OF THE STUDY

5.1 Introduction

Soybean is an important legume crop that is only ranked second to groundnuts as a reliable source of oil and protein. This crop is widely grown in different parts of southern Africa exhibiting different climatic conditions. Currently the main challenge of soybean crop is to develop improved varieties that will uniformly perform better than the available cultivars across their growing areas. Yield instability and lack of knowledge of genetic diversity are some of the factors contributing to the current low productivity. The objectives of the study were: 1) to determine yield stability and adaptability for elite soybean lines across six locations, 2) to understand genotype by trait associations of multiple trait relationships for the soybean elite lines across six locations and 3) to assess the level of genetic diversity among the elite soybean lines using molecular markers

5.2 Summary of results

5.2.1 Yield adaptation and stability analysis of tropical soybean

- Twenty-six elite soybean lines along with four checks were used to generate data used in this analysis during the 2016/2017 growing season in three countries and six locations. The data was subjected to AMMI and GGE biplot analysis.
- Soybean yield was significantly affected by genotype, environment and genotype by environment as revealed through the AMMI analysis.
- GGE biplot analysis using the “which won where” pattern identified genotypes with specific adaptation such as TGX 2002-4FM to environment E5.
- Genotypes TGX 2001-3FM and TGX 2001-16FM had general adaptation and high yield.
- TGX 2001-3FM was identified as an ideal genotype with high yield and highly stable hence it can be recommended for cultivar release in the tested environments.
- Environments E1 (Nampula) and E6 (Chitedze) were both in the same sector of the polygon view, hence one environment can be used for selection of the other environment to reduce the cost of the breeding trials.
- TGX 2001-15DM and TGX 2002-4FM had low yield mean performance but stable hence they can be further improved by using them as parents in another breeding pipeline by crossing to high yielding genotypes.

5.2.2 Correlation, path coefficient and genotype by trait association analysis among elite soybean lines across environments

- The GT biplot results, explained a high percentage of 80% of the total variation.
- Plant height had a positive direct effect on grain yield hence it can be used as a selection trait that contributes to yield that is, the more the values for plant height the higher the yield.
- TGX 2001-3FM, TGX 2002-3DM, and TGX 2002-6DM can be recommended to be used as parents in other breeding programmes for yield improvement.
- The positive relationship observed between grain yield and plant height; days to flowering and early vigour via days to maturity indicate the possibility of simultaneous improvement of the traits through selection.

5.2.3 Genetic diversity analysis of soybean

- The genetic diversity data was obtained using SNP markers from 48 soybean lines.
- The average gene diversity and genetic distance ranged from 0.42 to 0.55 with an average of 0.47 and 0.61 to 0.87 respectively.
- The polymorphic information content ranged from 0.44 to 0.5 with a mean of 0.48.
- Genotypes TGX 2002-3DM and TGX 2002-3FM had the highest genetic distance between them indicating that they were highly diverse.
- The AMOVA indicated highly significant differences at F=0.001 with among individuals, among populations and within individuals contributing 45%, 28% and 26% respectively.
- The 48 soybean lines were clustered in three main groups.
- From 348 SNP markers used 199 polymorphic SNPs genetically distinguished the genotypes. Therefore, the set of SNP markers used have the ability to accurately estimate genetic diversity among soybean genotypes used.
- The study revealed wide genetic diversity among soybean genotypes, therefore, creating an opportunity for parent selection that can be used for soybean improvement and increase productivity.

5.3 General recommendations based on the findings

The following breeding implications and recommendations realized from the study were as follows;

High yielding genotypes with general and specific adaptation were identified, informative environments and those with similar responses and were identified in the study. This entails that in the soybean breeding programme, when conducting multi location trials, the environments with the same response; one can be dropped to reduce the cost attached to carrying out the trials. The informative environments can be used as selection sites for genotypes prior from release in preliminary variety selection trials.

When selecting genotypes, both the grain yield and agronomic traits are very important hence the genotypes associated with high grain yield and their associated traits would be preferable to be used as female parents when making crosses. However, further studies can be performed on these lines to assess if these desirable traits can be passed on to the next generation.

Genotypes that were observed with wide genetic diversity can be used in a breeding programme for crop improvement and develop cultivars that have high grain and improved nutritional attributes. The polymorphic SNP markers can be recommended for use in different soybean diversity studies and breeding programmes. The genotypes observed with a high genetic distance can be commended and selected as parents for crosses in order to come up with superior cultivars