GENOMIC INVESTIGATIONS FOR CASSAVA BIOFORTIFICATION WITH PRO-VITAMIN A CAROTENOIDS IN KENYA

BY

WILFRED ABINCHA MAGANGI

(10874222)

THIS THESIS IS SUBMITTED TO THE UNIVERSITY OF GHANA, LEGON IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF DOCTOR OF PHILOSOPHY DEGREE IN PLANT BREEDING

WEST AFRICA CENTRE FOR CROP IMPROVEMENT

COLLEGE OF BASIC AND APPLIED SCIENCES

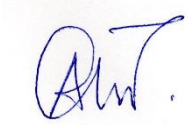
UNIVERSITY OF GHANA

LEGON

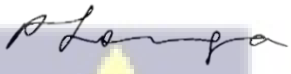
MARCH 2025

DECLARATION

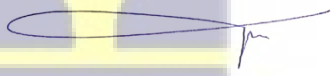
I hereby declare that this thesis is the product of my original research and has not, in whole or in part, been submitted for a degree in Ghana or anywhere else, with the exception of references to other peoples' works that have been properly cited.



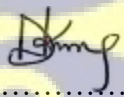
……..
WILFRED ABINCHA MAGANGI
(Student)



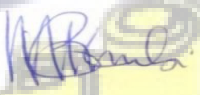
……..
PROF. PANGIRAYI TONGOONA
(Supervisor)



……
PROF. KWADWO OFORI
(Supervisor)



……
PROF. DANIEL KWADJO DZIDZIENYO
(Supervisor)



……..
DR. BENJAMIN MUSEMBI KIVUVA
(Supervisor)

ABSTRACT

Cassava (*Manihot esculenta* Crantz) is an important crop that serves as food and income source for about 800 million people in the world. The crop is popular due to its ability to grow in degraded soils, tolerance to extreme weather, requires less farm inputs to grow, and its ability to be harvested in piece meal when needed in small quantities. Despite cassava being rich in starch, it is deficient of other nutrients such as pro-vitamin A carotenoids. It has been noted that 68% of Kenyan children from communities that solely depend on cassava have vitamin A deficiency (VAD). To address this challenge, the Kenya Agricultural and Livestock Research Organization (KALRO) joined global efforts to biofortify cassava with pro-vitamin A carotenoids (pVAC). Genomic investigation was carried out to support cassava biofortification in Kenya. A population of 94 pVAC cassava genotypes were genotyped using SNP markers and phenotyped for beta-carotene content using high performance liquid chromatography (HPLC). Population structure, genetic diversity and linkage disequilibrium in the population was determined using R statistical software. Parametric genomic prediction models were developed using Bayesian generalized linear regression (BGLR) models. The prediction ability of these models was determined before and after adding significant SNPs from random marker effect Genome-wide Association Studies (GWAS) model into genomic prediction model. Similarly, non-parametric machine learning (ML) models for pVAC and their ensemble were developed and their prediction ability determined. In addition, this study investigated genomic regions, SNP marker superior alleles and candidate genes controlling pVAC in cassava roots using multi-locus and haplotype-based GWAS models. The results indicated that polymorphic information content (PIC) varied from 0.10 to 0.38, with an average of 0.24, implying that the markers used in this study were informative. The range of the minor allele frequency (MAF) was 0.05 to 0.50, with an average of 0.20. This indicates that the

genotype quality of the markers were reliable. Population structure results indicated that the population was structured in three ancestral sub-populations. According to statistics on genetic diversity, the observed heterozygosity (Ho) had a mean of 0.30 with a range of 0.21 to 0.38, while the genetic diversity (GD), also known as expected heterozygosity (He), had a mean of 0.29 and ranged between 0.10 and 0.50. This result revealed the presence of a moderate diversity in the study population. The coefficient of inbreeding (F) varied from -0.28 to 0.28, with an average of -0.02 and this indicated that the population was generally in random mating. According to analysis of molecular variance (AMOVA), 96.05% of genetic variability existed within the sub-populations while 3.95% existed between the sub-populations. The results showed that at an LD threshold of R^2=0.2, LD decayed at 613.072 kb while at R^2=0.1, LD decayed at 1786.714 kb. This result suggested that genomic analyses can be effectively carried out using a minimum of 1,256 markers distributed well across the genome. The addition of significant SNPs from GWAS enhanced the BGLR genomic prediction models. When more significant SNPs were included as fixed effects in the GP models, an improvement in the models was detected. Non-parametric ML models were shown to be accurate in predicting individuals' genomic value. ML model prediction power was greatly increased by feature selection and use of ensemble of models. The excellent prediction power of the Random Forest (RF), Extreme Gradient Boosting (XGBOOST), K-Nearest Neighbors (KNN), and ensemble of models makes them suitable for the last stage of choosing a potential cassava variety. The fixed marker effect GWAS model identified two significant SNPs while the random marker effect GWAS models identified five significant SNPs associated with beta-carotene. Haplotype based GWAS identified 15 genomic regions in 10 chromosomes associated with the trait. This indicated that among the studied GWAS models, haplotype-based GWAS had high statistical power. New causative variants on chromosomes 03, 05, 06, 08, 09, 10,

11 and 18 that are strongly associated with pVAC were discovered through this study. The investigation found alleles C, G, C, T, and G respectively on chromosomes 01, 03, 04, 09, and 14 to be superior alleles. These superior alleles leverage the need for quick biofortification of cassava with pVAC through gene pyramiding breeding strategy. This study uncovered a total of 20 candidate genes controlling carotenoid content in cassava roots. Five of the identified candidate genes (*Manes.01G124200*, Manes.01G001200, *Manes.05G193700, Manes.08G037100*, and *Manes.03G084700*) are involved in the anabolism of carotenoid content in cassava roots. This is the first time to report the detection of the latter three genes on the genomic regions associated with pVAC. On the other hand, the remaining genes (15) were involved in carotenoids catabolism. Identification of these candidate genes facilitates informed selection targeting the development of novel cassava varieties with high nutritional value. Results from this research directly contributes to the ongoing effort of biofortifying cassava with high pVAC at KALRO.

DEDICATION

I dedicate this thesis to my wife Dinah Rabera and my friends who consistently believed in me, encouraged and stood by me during my studies.

ACKNOWLEDGEMENTS

I give thanks to God for giving grace that has carried me through this work. I acknowledge the Director and staff of West Africa Centre for Crop Improvement (WACCI) whose efforts yielded for me to get WACCI-German Academic Exchange Service (DAAD) Scholarship. I also acknowledge the Director and Leader of Cassava Program of Kenya Agricultural and Livestock Research Organization (KALRO), Kakamega for giving me facilities to establish this research.

I acknowledge my research supervision team: Prof. Pangirayi Tongoona, Prof. Kwadwo Ofori, Dr. Daniel Dzidzienyo and Dr. Benjamin Musembi Kivuva for their guidance, and support during research and thesis development. I want also to acknowledge my colleagues and friends who shared their experiences and knowledge consequently helping me carry this research effectively. They include Dr. Ismail Siraj Kayondo, Dr. Alfred Ozimati and Dr. Nathaniel Ugochukwu Ikeougu.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

BGLR	Bayesian Generalized Linear Regression
CBSD	Cassava Brown Streak Disease
CIAT	International Centre for Tropical Agriculture
CMD	Cassava Mosaic Disease
DMC	Dry Matter Content
DAAD	German Academic Exchange Service
D.R.Congo	Democratic Republic of Congo
GCA	General Combining Ability
GEBV	Genomic Estimated Breeding Values
GLM	Generalized Linear Model
GP	Genomic Prediction
GWAS	Genome-Wide Association Studies
HCN	Hydrogen Cyanide
KALRO	Kenya Agricultural and Livestock Research Organization
KASP	Kompetitive Allele Specific Polymerase Chain Reaction
LD	Linkage Disequilibrium
ML-GWAS	Multi-locus Genome-Wide Association Studies
MLM	Mixed Linear Model
mrMLM	Multi-locus Random marker effect Mixed Linear Model
MT	Metric Tonnes
pVAC	pro-Vitamin A carotenoids
SL-GWAS	Single-locus Genome-Wide Association Studies
SNP	Single Nucleotide Polymorphism
TCC	Total Carotenoid Content
VAD	Vitamin A Deficiency
WACCI	West Africa Centre for Crop Improvement
WHO	World Health Organization

CHAPTER ONE

1.0 GENERAL INTRODUCTION

Cassava (*Manihot esculenta* Crantz) is an important crop that is mainly grown for food in sub-Saharan Africa. This crop serves as staple and income source for about 800 million people in tropics and subtropics (Bayata, 2019). Cassava is mainly cultivated for its roots that are highly rich in starch and in some regions, its leaves are consumed as vegetables. Cassava leaves are rich in proteins, vitamin B1, B2 and C, carotenoids and minerals (Latif & Müller, 2015). Fresh roots of cassava have energy production capacity of up to 610 kJ/100g (Temesgen et al., 2019).

Due to cassava's tolerance to a number of stresses, it has become popular, more so to resource-constrained farmers (Pushpalatha & Gangadharan, 2020). Compared to other food staples such as legumes and cereals, cassava is able to perform well in degraded soils. Cassava's long cropping cycle makes it possible to be harvested in piece-meal when only needed (El-Sharkawy, 1993; Howeler, 2017). All these attributes have made cassava popular in tropics and subtropics where it is grown as food security crop.

In Kenya, cassava is ranked as the second most important root crop after Irish potato (Githunguri & Gatheru, 2017). This crop is mainly grown in Western, Eastern and Coast regions of Kenya. In these regions, about 60% of the population depends on it as their staple (Mulu-Mutuku *et al*., 2013)

.

Unfortunately, cassava's role as a food security crop is constrained by its poor nutritional composition (Chavez et al., 2008). This property of cassava has negative impact on people who heavily consume it as their staple (Stephenson et al., 2010). Cassava roots are rich in starch but highly deficient in vitamins, fats and proteins (Ceballos et al., 2006). According to World Food

Summit of 1996, the concept of food secure world requires a world where at all times, people have physical, social and economic access to adequate, nutritious, culturally acceptable and safe food supply (FAO, 2006). Poor nutritional composition of cassava excludes it from the concept of food secure world. Among the nutrients that are lacking in cassava is vitamin A carotenoids. Vitamin A deficiency (VAD) is the major cause of night blindness (Lai et al., 2014). It has been noted that 250 million children are vulnerable to VAD, and of this, 32% is attributed to heavy reliance of cassava (FAO & WHO, 2001). In Kenya 68% of children in the regions that depend on cassava as their staple, suffer from VAD (Gegios et al., 2010).

Four strategies of addressing VAD have been suggested namely: fortification, supplementation, dietary diversification and biofortification. Fortification require people to buy fortified foods that are often expensive and out of reach to people in rural areas (Bhagwat et al., 2014). Similarly, food supplements are expensive and not readily available to majority in rural areas (Dickinson, 2014). Dietary diversification requires people to eat diverse food to achieve balanced diet (Lopez Villar, 2015). This creates a burden of growing different crops in remote areas which may not be successful due to different adaptation of crops in different locations. Deploying biofortified food relieves people the burden of buying expensive fortified food, supplements and diversifying foods which is only feasable to financially well off people staying in urban areas (Saltzman et al., 2013). Therefore, biofortification of staple crops with pVAC is the most effective strategy to combat VAD.

The perennial nature of cassava causes its genetic improvement to follow a long breeding scheme where a breeding cycle takes more than 8 years (Ceballos et al., 2016). It has been noted that selection based on genomic prediction (GP) facilitates the identification superior parents to cross at early stages of breeding, consequently, shortening the breeding cycle. However, to deploy the use of GP requires development of reliable genomic models. Despite the need to deploy GP in

Kenya for cassava biofortification, there is neither an established genomic model nor knowledge on strategies for achieving robust and reliable GP model. Indeed, in the whole cassava breeding program of KALRO, nothing has been done on the use of all available molecular signatures on cassava genome to predict cassava root quality traits. Presently, cassava breeders in Kenya use phenotypic recurrent selection scheme that is lengthy and involves expensive and laborious biochemical assays for quantification of traits such as carotenoids content. This has necessitated deployment of an alternative method such as GP that is able to increase genetic gains through reduction of selection cycle time and increase of selection accuracy and intensity.

Rigorous analysis on the genetic architecture of pVAC in cassava has not been achieved. Few studies that have been carried out to study the genetic architecture of the trait employed traditional GWAS using generalized linear model (GLM) and mixed linear model (MLM). However, it has been reported that the traditional GWAS models excludes known genomic regions associated with traits (Zhang et al., 2020). This is attributed to low statistical power and stringent threshold p-value typical of traditional GWAS. It was therefore postulated that not all loci associated with pVAC in cassava root have been uncovered.

Currently, GP is being deployed to accelerate genetic gains in plant breeding programmes. This technology uses all markers across the genome to accurately predict genomic estimated breeding value (GEBV) of an individual (Jannink et al., 2010). These genomic estimated breeding values are used for timely selection of the most promising parents. Deploying GP will increase the rate of genetic gain in cassava biofortification programme. Increased genetic gain facilitates quick development of improved varieties thereby responding quickly to nutritional requirement of cassava consumers. Phenotyping of carotenoids using reliable methods such as high performance

liquid chromatography and UV-spectrophotometer is expensive and laborious. Over time, GS will become cost effective as it reduces reliance on extensive phenotyping

By identifying all loci controlling pVAC in cassava roots, there will be better understanding of the molecular mechanisms underlying the pVAC accumulation process. This is vital in genetic improvement of cultivars and makes it easier to identify superior marker alleles and genes governing this trait. Identification of genomic variant associated with pVAC will facilitate the use of markers assisted selection (MAS) in cassava biofortification programme. Uncovering more superior alleles associated with the trait will enhance the efficiency in pyramid breeding. Furthermore, identification of superior alleles will allow targeted genetic modification of crops through the use of gene editing techniques. This will ensure genetic enhancement of the trait without undesirable genetic linkages.

The overall objective of this study was to support cassava biofortification with pro-vitamin A carotenoids in Kenya through development of molecular tools and accumulation of knowledge for efficient breeding.

The specific objectives were to:

1. determine genetic diversity, population structure and extent of linkage disequilibrium decay of pro-vitamin A carotenoids in a cassava population established in Kenya;
2. develop genomic prediction models for pro-vitamin A cassava breeding in Kenya; and
3. identify genomic regions, superior alleles and candidate genes regulating the accumulation of pro-vitamin A carotenoids cassava

CHAPTER TWO

2.0 LITERATURE REVIEW

2.1 Origin and biology of cassava

2.1.1 Origin of cassava

Cassava (*Manihot esculenta* Crantz) most probably originated from the wild species *Manihot esculenta* subspecies *flabellifolia* (Isendahl, 2011). Cassava has two significant geographical origins, one in Mexico and Central America and another in northeastern Brazil, according to archaeological data (Udoh et al., 2022). Cassava was transported from Brazil to the western coast of Africa by Portuguese navigators in the sixteenth century (Olsen & Schaal, 1999), and then to East Africa in the eighteenth century via the islands of Reunion, Madagascar, and Zanzibar (Hillocks et al., 2002).

The Portuguese established cassava plantations when they colonized South America around 1500 A.D. (Isendahl, 2011). Portuguese explorers brought cassava to Africa and Asia in the late 16th century by transporting it from these plantations to other continents. It spread across West and Central Africa after being first planted close to the Congo River basin (Olsen & Schaal, 1999). In the eighteenth century, Nigeria was one of the first African nations to acquire the crop. It was possible that freed slaves traveling back from South America via the islands of Sao Tome and Fernando Po brought cassava to southern Nigeria (Iwuagwu, 2012).

2.1.2 Biology of cassava

Cassava is a dicotyledonous, outcrossing plant possessing 18 chromosomes (2n = 36, n=18).

It is a perennial shrub with branching stems that range in height from 1 to 5 meters (Augusto & Alves, 2002). According to Fukuda et al. (2010), the palmate leaves are separated into three to

nine lobes. Male and female flowers are found in the same inflorescence, and the blooms are carried on auxiliary racemes close to the ends of branches. The fruit is a dehiscent, dry capsule with three seeds inside. Through a process known as secondary thickening, tubers-typically five to ten in number-develop radially around the base of the plant. The tubers are composed of a thin rind, or cortex, a starch-rich interior, or pith, and an outer skin, or periderm, which can be white, brown, or pink. According to Iwuagwu (2012), the core is often white, although it can also be yellow or reddish. The cylindrical stems of cassava range in diameter from 2 to 6 cm. Although cassava stems can reach a height of 4 meters, some genotypes may only reach 1 meter. Large, knob-like scars that resemble leaf scars and nodes can be seen on the older parts of the stems.

Cassava is grown primarily through stem cuttings, ensuring true-to-type cultivars. Nonetheless, multiplication by seed can occur naturally or during plant breeding activities. When stem cuttings are put in moist soil under suitable conditions, they sprout and form adventitious roots at the base of the cuttings within a week. When propagated from seeds, it first grows into a tap root system (Udoh et al., 2022).

2.2 Food security status in Kenya

With the projected 26 percent world population increase by 2050 (United Nations, 2019) , there is an urgent need to have reliable and increased food supply. Sub-Saharan Africa is expected to contribute for more than 50% of global population growth between 2022 and 2050 (UN, 2022). Reports are that as at 2023, a large portion of people; approximately 868 million in sub-Saharan Africa suffer a moderate food insecurity while about a third of this (342 million) suffer from severe food insecurity (FAO *et al.*, 2023). Earlier reports indicate that globally, about 24,000 people of whom 16,000 are children of less than 5 years die every day from hunger and malnutrition (UN, 2003). In 2022, According to FAO *et al*. (2023) Eastern Africa had the largest percentage (29%)

of undernourished people in Africa. In the same year, 38% of hunger stricken people in the world were resident in Africa.

Approximately 25% of Kenyan population (12 million people) have no access to sufficient food in terms of quantity and quality and therefore, rely on food aids throughout their life (Sibhatu et al., 2015). About 42% of pregnant women and 26% of children under the age of 5 years in Kenya have been reported to suffer from chronic malnutrition (MoALFC, 2020; USAID, 2017). Further, current reports have indicated that about 62% of Kenyan children suffer malnutrition attributed to vitamin A deficiency (Nutrition International, 2024).

Nutrition sensitive agriculture has been identified as a potential tool for addressing malnutrition issues (MoALFC, 2020). However, Kenya has a small portion of land that has potential for agricultural production since only 17% of Kenyan land is arable (Monke et al., 2019). Therefore, for Kenya to produce sufficient food, there is a need to use agricultural systems that will optimize yields and nutrition. From these reports, it is evident that food security status in Kenya is wanting and thus, there is a need for urgent intervention.

2.3 Economic importance of cassava

Cassava is a highly regarded food crop in Africa because of its ability to perform well in degraded soils, harsh environmental conditions and needs limited resources for cultivation. As a food security crop, both cassava roots and leaves are consumed. The leaves are consumed as vegetable and forage for animals (Latif & Müller, 2015). Apart from being a food security crop, cassava is now being produced as a cash generating crop. This crop is used in industrial production where the roots are processed into various forms such as chips, flour and starch. Starch extracted from

cassava has numerous industrial uses, which includes manufacture of products such as adhesives, maltose, fructose syrup, ethanol, thickeners and biodegradable plastics (Balagopalan, 2002).

Cassava has phytochemicals that possess medicinal properties such as remedy for cardiovascular diseases (Temesgen et al., 2019). Cassava has iron and copper which together play a role in the synthesis of red blood cells and this prevents anemia and other related diseases (Adeniji et al., 2010). With optimum red blood cells, body organs are supplied with enough oxygen which keeps cells and tissue at optimum health. The crop is rich in fiber that solves gastrointestinal tract problems and this eliminates bloating, constipation and even colorectal cancer (Temesgen et al., 2019). The crop is rich in potassium which is a vasodilator that reduces stress and tension of blood vessels. This reduces cardiovascular strain and minimizes chances of clots being stuck thereby reducing chances of stroke and heart attack (Adeniji et al., 2010). Other medicinal benefit from cassava include increased bone and neurological health, treating prostrate issues, allergies and diabetes (Temesgen et al., 2019).

Majority of rural communities in sub-Saharan Africa solely depend on cassava and consume it on a daily basis. These communities perceive that this is their food security crop and they have planted it in most household farms. Research has found that cassava is a suitable crop for rural farming and local food production systems (Saediman et al., 2016). Therefore, this crop should be improved to keep serving the dietary need of rural communities in sub-Saharan Africa.

Cassava is primarily grown for human consumption in coastal, western and eastern parts of Kenya where it serves as a main staple for the majority of the population. Due to cassava's resilience to water stress, the crop can be grown in arid and semi-arid parts of Kenya which comprises of over 80% of Kenyan land mass (Githunguri et al., 2015; Monke et al., 2019). Ability of cassava to

grow well using limited resources in regions where other crops cannot perform well makes it a food security crop in Kenya.

2.4 Focal traits for cassava improvement in Kenya

The most important part of cassava are the roots, which have low amounts of fats, carotenoids and protein content but also high amount of starch in the form of amylose and amylopectin content (Sanchez et al., 2009). Cassava roots have cyanogenic glucosides that are poisonous to human beings when roots with high levels are consumed. Therefore, cassava roots with high cyanogenic glucosides are processed prior to consumption to reduce the glucosides. Currently, researchers are focusing to improve cassava with high quality starch content, high DMC, high carotenoids content, reduced hydrogen cyanide content, enhanced yields, early maturing and disease resistance.

2.4.1 Root dry matter content and starch

Compared with other starchy crops, cassava yields the highest amount of starch per unit area (Tonukari, 2004). Cassava roots are a major starch sink and this starch possess reputable characteristics and comparatively, it is easier to extract because the roots have low levels of fats and proteins (Ceballos et al., 2007). Starch is estimated from DMC percentage because the two traits are highly correlated (r = 0.81) (Jennings & Iglesias, 2002). Indeed, the predominant portion of DMC in cassava roots is starch (Okechukwu & Dixon, 2008).

The quality of starch and its function is influenced by the amount of amylose content in cassava roots. Good cooking varieties have amylose content of 21%, varieties for industrial starch have amylose content of 15% and 17% for multi-purpose (Jennings & Iglesias, 2002). At CIAT, cassava breeders have been able to identify amylose-free starch mutant cassava which can be used for diverse industrial requirements (Ceballos et al., 2007). Therefore, in an effort to respond to various

starch demands, cassava breeders in Kenya are working towards achieving genotypes with desirable quality starch and DMC for domestic and industrial purposes (Saggafu et al., 2019).

2.4.2 Hydrogen cyanide level

All cassava plant organs except seeds contain some amount of cyanogenic glucosides (Augusto & Alves, 2002). Of these glucosides, linamarin is the most abundant (> 85%), while lotaustralin is in lesser amounts (<10%) (White et al., 1998). Linamarin is synthesized in the leaves and then transported to the roots where it is broken down by linamarase. Hydrogen cyanide (HCN); a volatile poison (LD50-60mg) for humans is produced when linamarin is hydrolyzed (Augusto & Alves, 2002). To reduce HCN levels to safe levels, one has to ferment, soak, heat, dry or combine all these treatments (Burns et al., 2010).

There is diversity in cassava genotypes based on HCN levels. Higher levels of HCN (180-720µg/g fresh weight) are found in root cortex and this is 2-6 times higher than the parenchyma (Piero et al., 2015). The level of HCN in cassava is influenced by environmental factors (Salvador et al., 2014). Sweet cassava genotypes have HCN levels less than 100mg kg^{-1} (100ppm) fresh weight while bitter cassava genotypes have HCN levels of 100-500mg kg^{-1} fresh weight (Augusto & Alves, 2002). Research has shown that in Kenya, majority of cassava in the market are unsafe for consumption with HCN levels ranging between 27.20 - 42.92 mg/kg, which is higher than 10mg/kg (10ppm) level as recommended by WHO (Gacheru et al., 2015; Gervason A et al., 2017). Nevertheless, efforts are underway through research to reduce cassava HCN levels to safe levels as recommended by WHO.

2.4.3 Carotenoid content

The main pro-vitamin A carotenoids in cassava roots is beta-carotene (Rodriguez-amaya & Kimura, 2004). Little levels of beta-carotene found in cassava roots can be reduced to retinol when

necessary, transformed into retinal when needed and stored in the liver after being esterified to fatty acids (Montagnac et al., 2009). Breeding for pVAC cassava in Kenya is a focal initiative that is expected to significantly support cassava consuming communities who have been reported to be vulnerable to VAD (Gegios et al., 2010). Previous effort to biofortify cassava with pVAC in Kenya aimed at germplasm collection and phenotypic characterization of the genetic resources (Njenga et al., 2022). The previous pVAC cassava-breeding scheme followed a lengthy phenotypic recurrent selection. However, the current pVAC biofortification effort in Kenya is directed towards the use of molecular tools to increase genetic gains by reducing selection cycle time.

2.4.4 Yield

Breeding for high cassava yields in Kenya is crucial for food security and improved livelihoods. Communities consuming cassava in Kenya have a significant reliance on the crop as a staple, especially in arid and semi-arid regions where other crops struggle to thrive. Over the years, cassava breeders in Kenya have been selecting high-yielding varieties adapted to local conditions. Regulations are set that every new variety must be high yielding compared to the present varieties. Yield related traits that cassava breeders focus in Kenya include root number per plant, root weight per plant and harvest index (Kamau et al., 2010). According to FAO (2024), cassava yield in Kenya stands at 11.78 MT/ha and 11.63 MT/ha in 2022 and 2023 respectively which is far below its potential of 80 MT/ha (FAO, 2013). Interestingly, it has been reported that some cassava varieties in Kenya such as MM99/0067 can yield as high as 200 MT/ha in one place but in most places it yields down to 9.4MT/ha (Wambua et al., 2020). Therefore, there is an effort directed toward production of improved cassava varieties with high yield in Kenya.

2.4.5 Resistance to diseases

Disease resistance is a focal trait for cassava breeding in Kenya due to the prevalence of devastating diseases like Cassava Mosaic Disease (CMD) and Cassava Brown Streak Disease (CBSD). These two diseases can cause up to 100% yield loss and thus, they are a threat to food security and livelihood of cassava consumers. Cassava breeders have prioritized disease resistance in cassava breeding programs to enhance yields and reduce the use of pesticides (Ntui et al., 2024). The Kenyan cassava breeding program are working to develop cassava varieties that have dual resistance against CMD and CBSD (Mutoni et al., 2021). Cassava breeding effort has yielded a number of cassava varieties that are tolerant to CMD and CBSD in Kenya (Musungayi et al., 2018). Furthermore, through genetic engineering, KALRO has developed cassava variety that is resistant to CBSD which has been approved for release (NBA, 2021).

2.4.6 Early bulking varieties

Cassava is a perennial crop with harvesting time ranging between 10 and 24 months after planting. Identification of early bulking varieties is one of focal objectives in Kenya's cassava breeding program. With recent climate change events leading to erratic rains and prolonged drought, identifying early bulking cassava varieties provide cassava farmers with ability to harvest before the adverse climatic conditions sets in. Through reduced harvesting time, early bulking cassava varieties reduce the risk of crop failure by ensuring timely harvest. In addition, early maturing cassava varieties promote diversified farming system by facilitating crop rotation or intercropping with other crops. Therefore, development of early bulking cassava varieties is one of the strategies for climate change adaptation.

Reports indicate that researchers at IITA, Ibadan, Nigeria, identified five cassava genotypes that relatively produced higher yields (25 to 28 tons per hectare) at six months after planting

(Okechukwu & Dixon, 2009). Researchers in Kenya have identified 10 cassava genotypes that are early bulking with high root quality harvested at 7 months after planting (Kamau et al., 2011). These genotypes were developed by hybridizing four popular Kenyan varieties with six early-bulking varieties selected from IITA germplasm.

2.5 Biochemistry and genetics of carotenoids in cassava roots

2.5.1 Biochemistry of carotenoids

The predominant type of carotenoids in cassava roots is beta-carotene (pro-vitamin A carotenoids) which is available in *trans* and *cis* isomers (Rodriguez-amaya & Kimura, 2004) (Figure 2.1). *Cis*-isomer has been found to have a lower vitamin A activity compared to *trans*-isomer (Rodriguez-amaya & Kimura, 2004). Bioconversion of pro-vitamin A carotenoids to vitamin A in human body is regulated naturally and thus carotenoids from biofortified crops do not have potential for toxicity compared to taking food fortified with vitamin A (Tanumihardjo et al., 2010). Vitamin A plays a fundamental role in human body of improving immunity, normal functioning of eyes, maintaining cellular integrity etc. (Combs & McClung, 2017). In cassava roots, carotenoids have been found to help in delaying post-harvest physiological deterioration (PPD) (Sanchez et al., 2006).

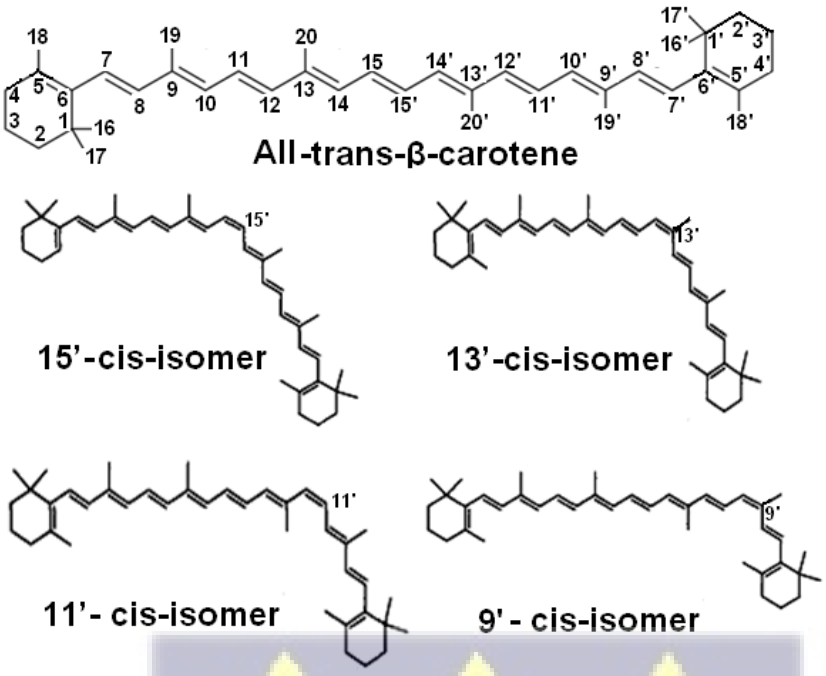


Figure 2. 1: Structure of trans-beta-carotene and isomers of cis-beta-carotene.

Carotenoids biosynthesis pathway shows the process of carotene biosynthesis in plants (Figure 2.2). The process of carotenoid biosynthesis in plants has been widely studied (Badejo, 2018; Udoh et al., 2022; Olayide et al., 2023) .The first reaction in the pathway is catalyzed by phytoene synthase (PSY) enzyme that condenses two geranylgeranyl pyrophosphate (GGPP) molecules into a colourless linear carotenoid compound called phytoene. Phytoene is modified into a red coloured carotenoid called lycopene through phytoene desaturase (PDS) and carotenoid isomerase (CRTISO) enzymes that catalyze this reaction. Lycopene synthesis stage is the forking point from where two separate downstream branches (α and β) forms. In α-branch, α-carotene and lutein are biosynthesized while β-carotene, zeaxanthin and β-crypoxanthin are synthesized in the β-branch through cyclization of lycopene's terminal. In the branched parts of the pathway, the main enzymes involved include lycopene epsilon α-cyclase (LYCε), lycopene epsilon β-cyclase (LYCβ) and β-carotene hydroxylase. LYCβ adds β-ionone rings to both ends of lycopene to produce β-carotene.

LYCε adds ε-ring in one end of lycopene to give α-carotene. Beta-carotene has full and high vitamin A activity because of its β-ionone rings on both ends. Hydroxylation at the C-3 position of each ring of β-carotene and α-carotene leads to production of zeaxanthin and lutein respectively. Zeaxanthin produces violaxanthin which in turn produces abscisic acid (ABA).

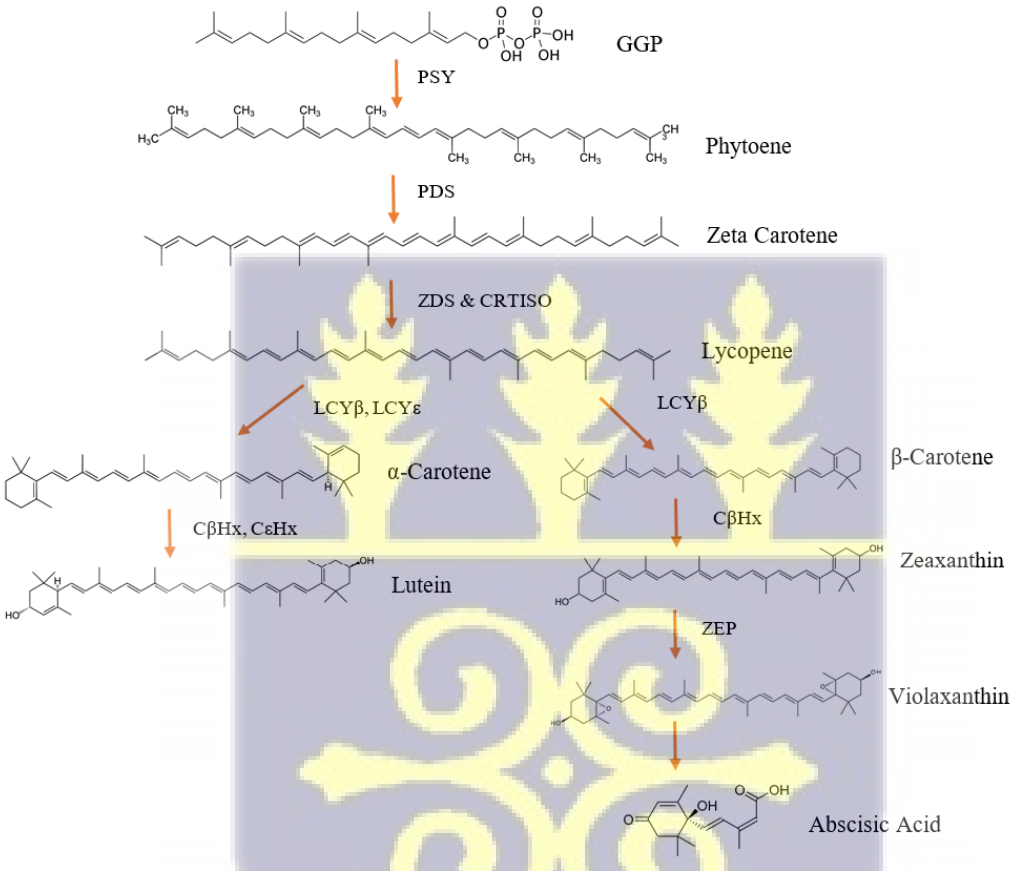


Figure 2. 2: Simplified carotenoids biosynthesis pathway. Adopted from (Udoh et al., 2022)

The carotenoid biosynthesis pathway helps breeders in understanding inheritance of carotenoids. To enhance carotenoids in cassava, breeders should halt conversion of β-carotene to zeaxanthin. This can be achieved through selection against β-carotene hydroxylase (CβHx) enzyme. There is no need for breeders to regulate the activity of LYCε because naturally the pathway favours

accumulation of β-carotene (Ceballos et al., 2017). Phytoene synthase (PSY) enzyme, which is responsible for biosynthesis of phytoene, occurs in three forms notably: PSY1, PSY2 and PSY3 (Arango et al., 2010). The amount of PSY3 is negligible, PSY2 is responsible for biosynthesis and accumulation of carotenoids while PSY1 is responsible for stress response through ABA (Ceballos et al., 2017). Therefore, breeders have to devote their effort to enhancing of PSY2 activity and deregulate PSY1 activities

2.5.2 Genetics of Carotenoids

Carotenoids in cassava roots have a high (71% - 86%) narrow sense heritability trait (Esuma et al., 2016b; Morillo et al., 2012; Parkes et al., 2020). Even though accumulation of carotenoids in cassava roots is governed by major gene, the quantitative variability observed in clone root color is a suggestion that there is a number of genes with small effects involved in the process of accumulation (Morillo et al., 2012). Early research done by Akinwale et al. (2010) on the inheritance of carotenoids in cassava found that inheritance of this trait does not involve maternal or cytoplasmic effects. Conversely, research done by Njenga et al. (2014) revealed that carotenoid inheritance in cassava roots is influenced by maternal and cytoplasm effects. This incongruent report from different researchers results from the fact that cassava is highly heterozygous and has no inbred lines that can be used to precisely study gene action for different cassava traits. This is one of the challenges that cassava breeders faces during genetic improvement of the crop.

It has been reported by Parkes et al. (2020) that carotenoids in cassava roots have a high significant ($P<0.001$) general combining ability (GCA) effects. This indicates that additive genetic effects control the inheritance of the trait and therefore, the performance of progenies can be predicted from the progenitors. Other related research by Esuma et al. (2016b) reports that selection for high carotenoid content in cassava roots can effectively be done in one location because the trait is not

significantly influenced by environment. This report is similar to the work done and reported by Ssemakula et al. (2007). In African cassava germplasm, carotenoids in cassava roots are negatively related with DMC (r = -0.42) (Akinwale et al., 2010). Similar observation was made in research done by (Esuma et al., 2016a). On the contrary, other studies have shown positive correlation in Latin American cassava germplasm (Ortiz et al., 2011). Recent studies indicate that this negative correlation between carotenoids and DMC in African cassava germplasm most probably could be due to linkage ruling out pleiotropy (Rabbi et al., 2017a). In cassava roots, carotenoid content has been found to be higher in proximal part of the root and gradually decrease in the central to the distal sections (Chavez et al., 2008). Research has revealed that carotenoids root-to-root variation within a cassava plant is greater than plant-to-plant variation (Ortiz et al., 2011).

2.6 Nutrient status of cassava roots and strategies for combating Vitamin A Deficiency in Kenya

Cassava roots have extremely small amount of proteins and lipids compared to other staple food such as cereals and legumes (Bayata, 2019). According to Gegios et al.(2010), other nutrients that are limited in cassava are zinc (about 3µg/100g), vitamin A (< 1µg/100g) and iron (about 4µg/100g). On the other hand, cassava leaves are highly enriched with vitamin A (8300µg/100g) and high amounts of zinc (71ppm) compared to the roots (Salvador et al., 2014). Consequently, it has been observed that communities in regions that depend on cassava as a sole staple are vulnerable to vitamin A deficiency. Sadly, 89% of children in Kenyan cassava growing region (western, eastern and coastal parts of Kenya) suffer from vitamin A deficiency (VAD) (Gegios *et al.*, 2010). Concerning this, effort has been put globally in international and national agricultural research institutions to biofortify this crop with micronutrients including pro-vitamin A carotenoid.

Vitamin A is a nutrient that is responsible for normal functioning of the eye; it is an antioxidant and thus boosts immunity (Combs & McClung, 2017). Furthermore, research has revealed that carotenoids delay post-harvest physiological deterioration (PPD) of cassava roots (Sanchez et al., 2006; Udogu et al., 2021). Several strategies have been suggested for combating Vitamin A deficiency, some of which are provided below.

2.6.1 Fortification

Fortification is the act of incorporating nutrients into food with the aim of improving overall nutrient quality of a diet. When fortifying a particular food, an appropriate 'food vehicle' is identified in which the target nutrient is added (Bhagwat et al., 2014). Among many target nutrients for fortification is vitamin A, iron and zinc (Das et al., 2013). In Kenya, maize flour, margarines have been fortified with vitamin A; this is tailored to reduce vitamin A deficiency. Sadly, it has been observed that fortified food is generally unreachable and expensive to the poor who usually are at a higher risk of vitamin A deficiency (Bhagwat et al., 2014; Makokha & Tunje, 2005). Considering this, fortification could not be effective in alleviating vitamin A deficiency among the poor majority. Furthermore, in some individuals, fortification has been found to be associated with the burden of diarrhea, intestinal diseases and mal-absorption; thereby limiting its efficiency (Das et al. 2013).

2.6.2 Dietary diversification

This strategy advocates for increased supply and consumption of food rich in micronutrients to achieve a balanced diet. This strategy involves making a wider food selection that is associated with high vitamin A content and other nutrients for purchase so that consumers prepare varied meals (Lopez Villar, 2015). In this strategy, there is a need to provide farmers with agricultural skills that will enable them grow nutrient rich crops. Furthermore, this strategy needs favorable

climatic condition to prevail in human settlements for growth of diverse food crops; this can be a challenge in most areas. For instance, dietary diversification cannot be practical in drought prone regions. In consideration of this, dietary diversification cannot be a good strategy to fight vitamin A deficiency among resource poor farmers that are resident in drought stricken regions.

2.6.3 Supplementation

According to United States Food, Drug and Cosmetic Act, a dietary supplement is any product that contains one or more ingredients which supplements the diet (Dickinson, 2014). These supplements include minerals, meal supplements, vitamins, sports nutrition products and other related products that can be taken by mouth as a liquid, capsule or pill. Dietary supplements are used to treat nutritional deficiencies. However, it has been reported that high supplement doses increases the risk of getting cancer (Martínez et al., 2012). This due to the fact that some supplements contain active substances that have strong biological effects to the body. For instance, high doses of antioxidants eliminate free radicals in the body; and thus, this interferes with body oxidation processes, which is essential for body defense. This attribute of supplements makes them unsafe for human health. Furthermore, many resource-constrained population in rural areas only depend on their staple and cannot access and afford these supplements. The limitations of fortification, dietary diversification, and supplementation have justified the need of another alternative, which is biofortification of staple food crops.

2.6.4 Biofortification

Biofortification has been found to be the most cost effective and sustainable means of delivering nutrients to resource poor population (Saltzman et al., 2013). Biofortification is the increment of nutrients into food crops through breeding. Food crops that already have preferred consumption and agronomic traits like high yields are targeted for biofortification. Biofortification is cost

effective because it only involves one-time investment in crop plant breeding to supply nutrients for a long period.

Biofortification is employed in three forms: agronomic, genetic engineering and conventional. Agronomic biofortification provides temporary micronutrients through fertilizers and/or foliar application of minerals (Wang et al., 2016; Zou et al., 2019). Conventional biofortification involves plant breeding strategies where parental lines that are rich in nutrients are crossed and evaluated for several generations until the desired level of nutrient is achieved. Biofortication using genetic engineering platform employs either transgenic or gene editing technologies to improve crops and this gives higher gains than that of conventional breeding (Liu et al., 2021). Transgenic biofortification needs legal approval for it to be implemented and therefore conventional biofortification is the most popular.

2.7 Breeding for pro-vitamin A carotenoids cassava

Cassava is a starchy crop that is cultivated in the tropics for food and commercial use. Primary traits for cassava selection are DMC (Okechukwu & Dixon, 2008), and micronutrients particularly pro-vitamin A carotenoids. Inherently, cassava roots are starch sink, thus, cassava in developing world has been targeted for biofortification with pro-vitamin A carotenoids. Majority of cassava varieties grown across the world are white rooted with a few micronutrients (Welsch et al., 2010).

Biofortification of cassava with increased pro-vitamin A carotenoids is a reliable means of supplying nutrients to communities that depend on cassava as their staple (Saltzman et al., 2013). Biofortification of cassava will have a positive impact on nutrition to low income communities that cannot afford dietary diversification, fortified food and supplements. Efforts to biofortify cassava with pro-vitamin A carotenoids is underway in both international and national agricultural research systems. Global projects such as HarvestPlus are actively involved in cassava

biofortification so as to at least reach the daily requirement of vitamin A for children less than 5 years (400µg retinal activity equivalents).

In Kenya, cassava biofortification with pro-vitamin A carotenoids is in initial stages. Breeding of cassava in Kenya is predominantly phenotypic recurrent selection (Abincha et al., 2024; Njenga et al., 2014). Previously, Njenga et al.(2010) reported introduction of pro-vitamin A carotenoids cassava population in Kenya. This population was phenotypically characterized on its variability for use in cassava biofortification (Njenga et al., 2018). Unfortunately, breeding efforts for pro-vitamin A cassava stalled at KALRO following the loss of the foundational plant population due to conservation problems.

2.7.1 Cassava breeding scheme

Cassava breeding follows a phenotypic recurrent selection but unlike other crops, it is not efficient due to presence of constraints (Ceballos et al., 2004). Recurrent selection is a breeding technique in which individuals are chosen, and then they interbreed to raise the frequency of alleles that are desirable in the population or lower unfavorable alleles(Singh et al., 2022). Phenotypic recurrent selection involves phenotyping individuals once before selection, which serves as the sole evaluation of the individuals' breeding value (Labroo & Rutkoski, 2022). Elite cassava genotypes are crossed resulting in segregating and highly heterozygous progenies. From the progenies, superior genotypes are identified and vegetatively multiplied. This recurrent selection has some similarity with hybrid crop breeding. However, there is huge differences because cassava has no inbred lines and therefore, there is no proper separation between general (GCA) and specific (SCA) combining abilities. Pollination is slow and thus, it takes about 2 years (18-24 months) from planning of crosses to obtaining of botanical seeds. The botanical seeds are grown in the third year to establish a seedling population. The seedlings are cloned in clonal evaluation trials (CET) and

planted in the fourth year for clonal evaluation. Therefore, typical cassava phenotypic recurrent selection does not take less than four years (From planning of the cross to first evaluation in a CET). However, this duration can be longer if selfing of the population is carried out to reduce genetic load.

Cassava breeding scheme is long since the breeding cycle takes more than 8 years (Ceballos et al., 2016). Outstanding parental genotypes are selected and botanical seed obtained through crossing. Obtained seeds are germinated in a greenhouse and F_1 seedlings transplanted into the field after about two months. Selection is done at 9-10 months after planting and the only selection criteria is the ability to produce 8 vegetative cuttings of 20cm each.

In subsequent selection, cassava breeders face a wide range of objectives due to great array of cassava production environments, crop management and end use preferences. However, only a few broadly accepted traits are considered for improvements; and these include fresh root yield (FRY), total carotenoids content (TCC), dry matter content (DMC), resistance to diseases and pests, and plant architecture. Due to the low heritability of FRY in the early phases of selection, cassava breeders have used correlated traits with greater heritabilities, including harvest index (HI), to apply indirect selection for yield for many years (Ojulong et al., 2010). A selection index (SI) which integrates all these important traits is used for selection (Eqn.1). In the SI, each trait is assigned weight based on the breeder's judgement.

$$SI = (TCC * 10) + (DMC * 10) + (FRY * 10) + (HI * 5) - (CBSD * 4) \quad \text{.....Eqn.1}$$

2.7.1.1 Clonal evaluation trials (CET) or single row trials (SRT)

This is the initial stage where selection of agronomic trait performance takes place. In this trial, cassava genotypes are planted in augmented design with single row plots of about 6-8 plants in a

single location. At 12 months after planting, genotypes with good agronomic traits are selected and advanced into the next evaluation stage (Figure 2.4).

2.7.1.2 Preliminary yield trials (PYT)

At this stage of evaluation, each genotype is planted in a plot of 10 plants (two rows each with 5 plants). Planting is carried out using incomplete bock design or randomized complete block design with three replications. Plants in PYT are planted in a single location and harvesting is done on all plants except the front plant in each row.

2.7.1.3 Advanced yield trials (AYT)

Genotypes are planted in plots of 4-5 rows, each row containing five plants. This is planted in incomplete block design or randomized complete block design with three replications. Six to nine central plants in a plot are harvested to generate data for evaluation. AYT are planted in a single location.

2.7.1.4 Uniform yield trials (UYT)

This is the final stage of evaluation and the plot size, number of replications are same as those in AYT and PYT. Evaluation takes place in two years across four to ten locations. UYT usually has 20-25 clones to be evaluated and about 8 checks. During harvesting, cassava users are invited to participate in selection and evaluation process.

This cassava-breeding scheme can be modified to fit the breeder's objective. For instance, Kawuki et al.(2011) modified this scheme by setting PYT across locations, removing AYT and setting UYT with many genotypes. This modification was done in order release improved cassava varieties within a period of 5 years

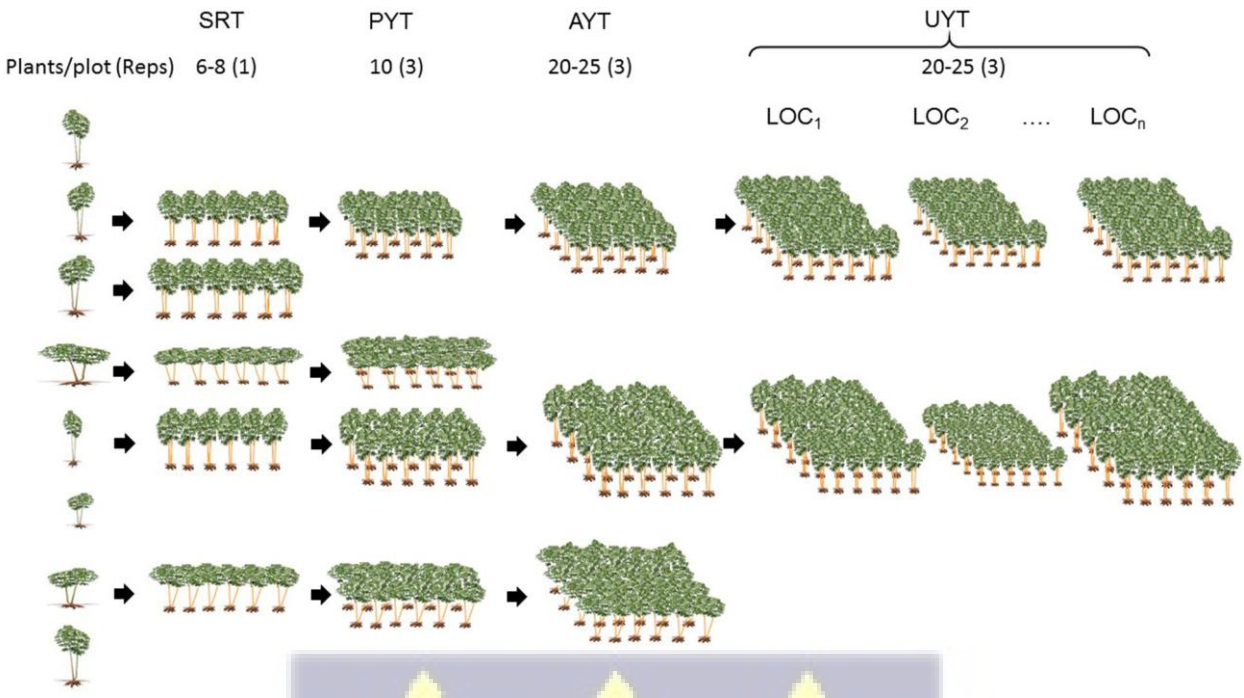


Figure 2. 3: Cassava breeding scheme at CIAT. Source: Ceballos et al.(2016)

2.7.2 Rapid cycling recurrent selection

Rapid cycling recurrent selection cycling recurrent selection is a breeding scheme devised only for pVAC in cassava roots. This breeding strategy shortens the breeding cycle from 8 years taken by phenotypic recurrent selection to 3 years. For several years, research has revealed that carotenoid content in cassava roots is a high heritability trait and this has been recently demonstrated in the work done by Esuma et al. (2016a); Morillo et al.(2012). Based on this, a fundamental modification of cassava breeding scheme was drafted at International Centre for Tropical Agriculture (CIAT) for genetic improvement of carotenoids in cassava roots (Ceballos et al., 2013). This new scheme is the rapid cycling recurrent selection that shortens the cycle of selection (Figure 2.4).

The rapid cycling recurrent selection for carotenoids improvement in cassava roots is achieved by selection of genotypes based on root pVAC levels at the seedling stage without evaluations across

locations. Selection of parents takes place at the end of F_1 trials (seedling stage) at 12 months after planting (MAP). In the second year of the cycle, selected parental genotypes are cloned and left in the field for 18 months. Here, crosses are made from 6-15 MAP since cassava has varied flowering habits. Fruits from early crosses are harvested and their seeds planted in the third year of the cycle. However, most of the fruits are planted in seedling trial in the fourth year of the cycle. Here, seedlings are transplanted into the field and the seedling plants evaluated based on their pVAC content. This kind of selection takes advantage of high heritability of carotenoid content in cassava

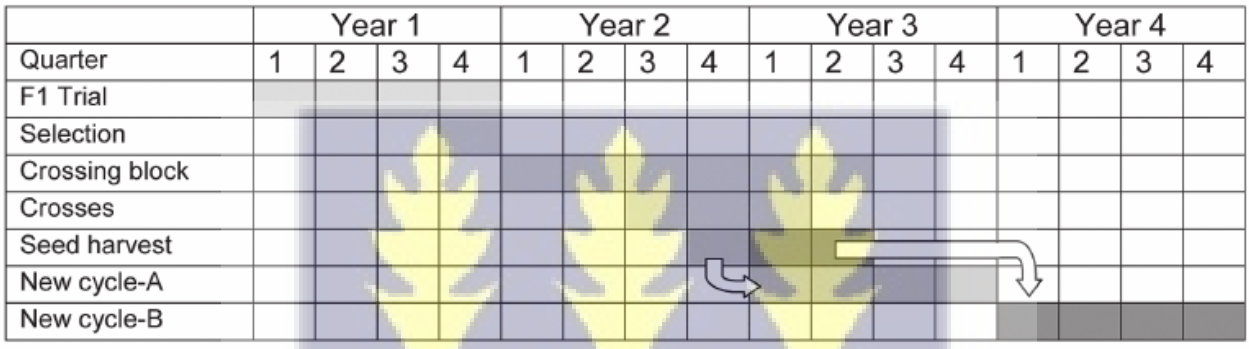


Figure 2. 4: Rapid cycling recurrent selection scheme. Source: Ceballos et al., (2013)

roots and therefore, it is not suitable for improving low heritability traits such as yield.

2.7.3 Genetic diversity in Kenya's pro-vitamin A carotenoids cassava population

Success in any breeding program is anchored on high selection intensity, selection accuracy, genetic diversity and reduced cycle time. Application of breeding tools that accelerate genetic gains over time such as genomic prediction and marker assisted breeding relies on the extent of genetic diversity (Jannink et al., 2010). Furthermore, with continuous emergence of new stresses, there is need for diverse populations from which to source new alleles that will help in crop improvement with new traits.

The world cassava gene pool has natural variation of roots rich in carotenoids (Sánchez et al., 2014). This variation can be exploited to biofortify cassava with sufficient pro-vitamin A carotenoids. Several studies on genetic diversity of pro-vitamin A carotenoids cassava have been done, however, in east Africa, little has been done. Notable of these include the assessment of genetic diversity in Ugandan's core collection of pro-vitamin A cassava population that included Ugandan landraces, accession from CIAT and IITA (Esuma et al., 2012).

Genetic diversity studies targeting pro-vitamin A carotenoids in a cassava population in Kenya have not been carried out. It has been previously reported that pro-vitamin A cassava populations in Kenya consist of a few IITA genotypes hybridized with local white fleshed cassava cultivars (Njenga, et al., 2014). Unfortunately, this reported population got lost due to crop conservation challenges. Only a few cassava varieties conserved at IITA Kenya are available. Concerning this, the Kenyan cassava breeding program has collected new pro-vitamin A carotenoids cassava germplasm on which rigorous characterization of its diversity is underway. This will facilitate cassava breeding program in Kenya with reliable germplasm and an opportunity to make informed decisions in management, establishing conservation and breeding strategies for increased carotenoids in cassava roots.

2.8 Pro-vitamin A carotenoids cassava breeding tools

2.8.1 Genomics

The use of genomic prediction in pVAC cassava breeding is employed to address the challenge associated with the crop's long breeding cycle. Genomic prediction models use all marker effects in cassava genome to predict a trait of interest. The model is applied to estimate genomic breeding values (GEBVs), which enable early identification and selection of superior parental lines. Genomic prediction models increase the genetic gain through reduced cycle time, increased

selection accuracy and intensity. Reports indicate promising results on the use of genomic prediction models for accelerated cassava breeding with pVAC (Esuma et al., 2021; Ikeogu et al., 2019). It is worth noting that genomic selection predicts breeding values and not genotypic value since it doesn't take account of non-additive genetic effects.

Genetic architecture of pVAC and correlated traits is determined through genome-wide association studies (GWAS). From this analysis, genomic signatures associated with the traits are identified and used to develop Kompetitive Allele Specific PCR (KASP) markers to use in marker assisted selection (MAS). Furthermore, results from GWAS can further be analyzed to identify candidate genes for the traits of interest. Previous GWAS studies of pVAC identified SNPS in chromosomes 1, 2, 4, 13 ,14 and 15 associated with different carotenoid isomers (Ikeogu et al., 2019).

Genomics is essential for studying cassava diversity, enabling detailed analysis of genetic variation among different cassava varieties. Through genotyping platforms, researchers use genetic markers and variations to reveal insights into cassava's genetic diversity, evolutionary history, and adaptation mechanisms. Information gained from the studies aids in preserving genetic resources and selecting promising genotypes for breeding programs. In Kenya, the use of genomic knowledge will support cassava breeding programs in enhancing cassava's nutritional quality, resilience and productivity, ultimately benefiting cassava farmers and enhancing food security.

2.8.2 Genome editing

Majority of genome editing technology are designed to make changes in the genome with high precision to create variation without introducing new genome to the organism (Gaj et al., 2016). Being a new technology, a few cases have been reported on its use on genetic improvement of cassava. The first report on cassava genome editing was on the use of CRISPR/Cas9 to knock down phytoene desaturase (PDS) gene (Odipio et al., 2017).

According to Srinivasan et al. (2017), PDS is an enzyme in the carotenoid biosynthesis pathway that transforms colorless phytoene into lycopene, a colored molecule. Therefore, albino phenotypes are typically produced when PDS is knocked out(Srinivasan et al., 2017). The majority of the plants had an albino phenotype when Odipio et al. (2017) altered PDS in the model cultivars TMS60444 and TME 204, resulting in a mutation frequency of 90% to 100%. Their findings paved the way for cassava genome editing to confer resistance against CMD and CBSD.

Following the successful use of genome editing technology to knock down PDS gene, researchers have been able to use the technology in developing cassava brown streak disease (CBSD) resistant varieties. Some host genes known as susceptibility genes (S genes) in plants are thought to be necessary for compatible plant-pathogen interactions because they facilitate pathogen invasion (Zaidi et al., 2018). The pathogen activates these genes during pathogen invasion to favor pathogen growth and promote symptom development, and editing S genes has been shown to confer resistance to the corresponding pathogen and, in some cases, broad-spectrum resistance (Jhu et al., 2023). Cassava plants produced by editing the susceptibility gene eIF4E (nCBP-1 and nCBP-2) showed partial resistance against CBSD (Gomez et al., 2019).

Similarly, two genes that produce the primary cyanogens in cassava, CYP79D1 and CYP79D2, were silenced by plant scientists using CRISPR/Cas9 editing at the Innovative Genomics Institute (IGI), a nonprofit research organization established through a collaboration with the University of Berkeley in California and University of California-San Francisco. With major advantages over conventional breeding techniques, the scientists pointed out that the application of CRISPR/Cas9 is a logical next step for breeding improvements into cassava. This is because the technology has

the ability to either mute the CYP79D genes separately or concurrently. The new plants are similar to what can be accomplished by conventional breeding, but in less time.

This technology can quickly address complex challenges affecting cassava breeding for pVAC. Among challenges that can be addressed by this technology is CBSD that causes up to 100% root damage ultimately affects cassava biofortification efforts (Gomez et al., 2019). This technology can also be used to quickly provide solution to the negative association of carotenoid content and DMC challenge in pVAC cassava.

2.8.3 Transcriptomics

Transcriptomics involves the study of genome-wide RNA transcripts and this is pivotal in cassava breeding. Transcriptomics is devoted to the analysis of gene expression profiles, through which researchers identify genes involved in desirable traits such as disease resistance, drought tolerance, and high pro-vitamin A carotenoid content. The detailed understanding of gene activity under differential conditions such as different environments, developmental stages, genotypes with contrasting trait levels is helpful in pinpointing genes that are activated or suppressed in response to specific condition. Furthermore, transcriptomic analysis also facilitates the development of molecular markers for marker-assisted selection (MAS) consequently, speeding up the breeding process.

Few transcriptomic reports on pVAC cassava have been established. Through transcriptomic studies, genes associated with pVAC accumulation in cassava roots have been identified (Olayide et al., 2023). Similarly, transcriptomic studies revealed that despite yellow and white fleshed cassava varieties not having significantly difference in carotenoids biosynthesis genes, yellow fleshed varieties have higher expression of lycopene-ε-cyclase (LCYε), phytoene synthase 2 (PSY2), and β-carotenoid hydroxylase (CHYβ) expression (Olayide et al., 2020). Other

transcriptomic studies have revealed that genes associated with myo-inositol and cell-wall polymer biosynthesis are significantly enriched in yellow fleshed cassava varieties (Gutschker et al., 2024).

2.8.4 Phenomics

Carotenoid phenotyping is crucial for measuring and quantifying the overall amount of carotene in cassava roots. Since some methods of estimating the total carotenoid content, including high-performance liquid chromatography (HPLC), can be very costly, selection becomes a drawback as breeding for higher carotenoid levels in cassava progresses. Additionally, it has been noted that the amount of carotenoids in cassava roots is directly correlated with the color intensity of the roots (Rabbi et al., 2017). Although visual selection is helpful in differentiating between cassava with white and yellow roots, it is ineffective in identifying the key distinctions between yellow roots. Additional techniques for measuring carotenoids or assessing color intensity include iCheck Carotene, Chromameter, ultraviolent-visible spectrophotometer, and near-infrared spectroscopy (NIRS).

2.8.4.1 Near-infrared Spectroscopy

Near-infrared Spectroscopy (NIRS) is a phenotyping method that can be utilized in the field right away and is offered in desk-top and portable forms (Abincha et al., 2020). NIRS screens samples in many states, such as solid and liquid forms, and enables the prompt screening of numerous samples and variables. In contrast to conventional phenotyping techniques, it offers a quick and nondestructive way to analyze multiple constituents at once with little to no sample preparation. It is cost-effective and poses no environmental risks (Abincha et al., 2021). According to Ikeogu et al. (2017), NIRS is inexpensive, versatile, precise, very portable, and fast at phenotyping hundreds of samples every day. The creation of a prediction model linking spectra and phenotype is necessary for the application of NIRS. According to recent findings, NIRS models created using

straightforward sampling reference alternatives can be used with confidence for cassava pVAC and DMC phenotyping (Abincha et al., 2021).

2.8.4.2 Chromameter

In cassava breeding programs, chromameters have been employed as an objective pre-selection tool for carotenoids based on the sample's color intensity (Sánchez et al., 2014). Chroma-meters use a xenon arc lamp to illuminate the sliced sample with a pulse or flood of light of a certain color or brightness in order to quantify the surface color and surface darkness/lightness to an exact degree. According to Abincha et al. (2020), Chromameter uses the Lab color space, which mathematically depicts all three-dimensional perceptible colors: L*, a*, and b*. L* stands for lightness, a* for green-red, and b* for blue-yellow. Typically, L*, a*, and b* have absolute values. The brightest white is represented by L* = 100, whereas the deepest black is represented by L* < 0. Neutral gray is represented by the variables L* = 0, a* = 0, and b* = 0. Green is represented by the a* axis's negative values, and red by its positive ones. Blue is represented by the b* axis's negative values, and yellow is represented by its positive values. The a* and b* axes' limits typically fall between -128 and +127 or within the range of ±100. Carotenoids in cassava roots are measured using positive values on the b* axis. On this axis, the intensity of yellowness rises as b* values shift from 0 to +100. Consequently, the positive b* axis values increase with the amount of carotenoids in cassava root.

In certain plant tissues, this straightforward method has also been applied to precisely measure color intensity and quality (Olayide et al., 2020). Sánchez et al. (2014) found a strong and positive correlation between total carotenoid content and color intensity ($R^2 = 0.769$). This suggests that a visual inspection of the parenchyma's color intensity can be used to select cassava clones with a relatively high total carotenoid content. Using the chromameter's L* a* b* color coordinate

method, de Carvalho et al.(2022) examined the color differences of 228 biofortified cassava clones. The results showed a strong positive association between total carotenoids content (TCC) and the chromameter's b* values (r = 0.90). Their findings show that using data from this device is a cost-effective, quick, and efficient method for TCC phenotyping.

2.8.4.3 iCheck Carotene

This is a portable equipment that is made up of two parts: the disposable reagent vial (iEx™) where the reaction is carried out and the measuring unit (iCheck™ Carotene). Two millilitres of a reagent combination, required to complete the reaction, are contained in the disposable reagent vial. With a weight of about 250 g, the iCheck Carotene is incredibly portable. This device has rechargeable batteries that makes it able to record up to 400 measurements, which are immediately saved and accessible via a USB cable as a text file whenever needed (Udoh et al., 2022). The iCheck Carotene is a quick screening technique that is affordable, easy to use, and economical. This device doesn't require highly qualified and specialized staff to operate, nor does it require a costly laboratory setup with specific chemicals and equipment (Maroya et al., 2012). The iCheck Carotene can be used to quickly and accurately quantify a large number of samples, particularly in situations where laboratories are unavailable.

2.8.4.4 Ultraviolet-visible (UV–vis) spectrophotometer

The basis for the use of spectrophotometry in the quantification of carotenoids is the idea that light is a type of electromagnetic radiation, and as such, when it strikes a substance, it can reflect, absorb, or absorb certain wavelengths of light while transmitting or reflecting the remaining wavelengths. Carotenoids in solution obey the Beer-Lambert law, whereby their absorbance of light is directly proportional to their concentration in solution. Thus, carotenoids are quantified spectrophotometrically, provided that accurate absorption coefficients in the desired solvent are

available (Rodriguez-amaya & Kimura, 2004). The limitation of this method is that it is laborious and unable to quantify the specific type of carotenoid in a sample.

The general procedure for spectrophotometry carotenoid quantification include sampling, extraction, partitioning, saponification and quantification. Sampling is important because the sample subjected to analysis must be representative of the lot under investigation. The sample must be appropriate in size and homogenized. Carotenoids are lipophilic, thus insoluble in water and soluble in organic solvents. Therefore, organic solvents i.e. acetone, methanol with low boiling points (35-60 degrees Celsius) should be used in carotenoids extraction to avoid overheating during evaporation. The extract usually contains a substantial amount of water, which can be removed by partitioning.

Saponification is an effective means of removing chlorophyll and unwanted lipids. However, cassava does not contain esterified carotenoids and has a low lipid content; hence saponification is unnecessary. The extracted carotenoid have different absorption wavelength depending on the type of solvent used during extraction. Total carotenoid is calculated using the formula:

$$TCC\ (\mu g/g) = \frac{\text{A x V(mL) x } 10^4}{A_{1cm}^{1\%}\ \text{x P(g)}}$$

Where A = absorbance; V = total extract volume; P = sample weight; $A_{1cm}^{1\%}$ =2592 (ß carotene extinction coefficient in solvent).

This method has been used in several studies notable of which include Chavez et al., 2008; Esuma et al., 2012; Morillo et al., 2012.

2.8.4.5 High Performance Liquid Chromatography

This is a sophisticated type of liquid chromatography that is used to separate, identify, and quantify the constituents of a mixture of molecules found in biological and chemical systems. It has a high rate of recovery, is easy to pick and manipulate, and is highly reproducible (Jaramillo et al., 2018). In its basic operation, a liquid (mobile phase) is pumped at high pressure into a column of porous material (stationary phase) while a solution of the sample is injected into the column. The partitioning of the sample between the stationary and mobile phases causes the sample to segregate according to the variations in migration rates through the column (Belalcazar et al., 2016; Jaramillo et al., 2018).

In cassava phenotyping, HPLC is utilized to separate and quantify individual carotenoids with varying pro-vitamin A activity. Although it has great reproducibility, its analysis is expensive, costing between $50 and $70 per sample with a relatively limited throughput (Abincha et al., 2020). It is time-consuming, labor-intensive, and necessitates a sophisticated laboratory setup with highly qualified workers and a tight quality control regimen (Jaramillo et al., 2018).

CHAPTER THREE

3.0 GENETIC DIVERSITY, POPULATION STRUCTURE AND LINKAGE DISEQUILIBRIUM OF A PRO-VITAMIN A CAROTENOIDS CASSAVA POPULATION IN KENYA

3.1 Introduction

The low nutritional content of cassava (*Manihot esculenta* Cranz) makes it unrecognizable as a whole food crop. People that mostly depend on cassava as a main food have been found to be at high risk of vitamin A deficiency (VAD). Notably, children in cassava consuming communities in Kenya were found to be highly vulnerable to VAD (Talsma et al., 2016). Evidence suggest that VAD leads to xerophthalmia, a general term for all ocular diseases resulting from night blindness due to corneal destruction (keratomalacia) (Belete et al., 2019).

Global efforts are directed towards development of pro-vitamin A carotenoids (pVAC) rich staple such as cassava (Foley et al., 2021). Kenya has joined this effort by initiating a cassava biofortification program. However, there is limited achievement in breeding and distribution of pVAC cassava varieties in the country. This could be attributed to lack of adequate and properly characterized pVAC cassava genetic resources. During plant variety development, plant-breeding programs develop strategies that lead to high genetic gains from breeding. These strategies depend on an understanding of genetic characteristics of available genetic resources. Therefore, a thorough understanding of the genetic characteristics present in Kenya pVAC cassava germplasm collections is necessary for successful utilization of the genetic resources in breeding programs. These genetic characteristics include genetic diversity, population structure, the extent of linkage disequilibrium (LD) and haplotype presence within the plant genetic resources.

This research aims at determining the genetic diversity, population structure and extent of LD within the assembled pVAC cassava germplasm.

3.2 Materials and Methods

3.2.1 Plant Material

Ninety-three (93) pVAC cassava genotypes were used in this study. These genotypes are half sib progenies of pVAC parents from cycle 2 (C_2) of selection accessed from National Crops Resources Research Institute (NaCRRI), Uganda. These parents were partial inbred lines developed through inbreeding and selected based on their carotenoid content. The seeds of genotypes were germinated at KALRO, Kakamega, Kenya from June-December 2022. Plastic germination bags were filled with sterilized forest soils and cassava seeds planted in them at a depth of 1.5cm. The bags together with the planted seeds were arranged in wooden crates that were covered with black polythene papers to retain heat. The screen house was fitted with heaters where temperature was adjusted to range between 30^0C- 40^0C as suggested by Mezzalira et al.(2013). Watering and monitoring was done daily and germinated seedlings collected for hardening. Hardening was done by first, removing germinated seedlings from the warm covered crates to open crates placed in a screen house. The seedlings were allowed to stay in the screen house where watering was carried out once after two days for one week. The seedlings were then placed into a net house for two weeks where they were watered once after three days. The hardened seedlings were transplanted into the field after having stayed in the net house for two weeks.

3.2.2 Genotyping

Cassava leaf samples were collected from a single plant in January 2024. Young leaves were chosen and punched using a leaf puncher to achieve uniform sample size. The puncher was well wiped with sterilized cotton using 70% ethanol before punching a new leaf sample. The samples

were placed into 96 well sample plate and oven dried at temperature of 40^0C at KALRO, Kakamega, Kenya laboratories. The dried leaf samples were shipped to SEQART Africa at ILRI, Nairobi, Kenya for DArTseq (GBS) (1.0) genotyping. SNP calling was carried out using *Manihot esculenta*_671_v8.0 reference genome and this yielded 54,877 SNPs.

3.2.3 Quality control

SNP quality control was achieved using *snpReady* package in R statistical software (R). The SNPs were filtered based on a call rate of 0.95, MAF < 0.05 and a sweep sample of 0.40. Based on this quality control, 37,419 SNPs were retained for further analysis.

3.2.4 Analysis

3.2.4.1 Population diversity and Structure

The SNP density plot was developed using *rMVP* package in R. Structure analysis was carried out using LEA package in R. Genetic diversity of the population was carried out using *snpReady* package in R. The diversity parameters determined were observed heterozygosity (Ho), expected heterozygosity (He), inbreeding coefficient (F_{IS}) and polymorphic information content (PIC) of the SNP markers.

To enhance the reliability of population structure analysis, SNPs that were in high LD level (r^2 => 0.70) across the genome were pruned using *ASRgenomics* package in R. Optimum number of ancestral populations (K) was identified using *snmf* function in R, which estimates admixture coefficients using sparse Non-Negative Matrix Factorization (sNMF) to provide STRUCTURE-like output. This analysis was run with ten ancestral populations (K=10), 10 repetitions for each K value and cross-entropy criterion. The minimal cross-entropy in function of K was plotted using *plot* function of *LEA* to identify the optimal number of ancestral populations. Ancestry coefficients

(Q matrix) were developed for the second run of K and its barplot plotted to display admixture proportions on the genotypes. The optimum number of populations were visualized through a dendrogram plotted using *dendextend*, *adegenet*, *NAM* and *circlize* packages in R. Analysis of molecular variance (AMOVA) was carried out using *poppr* package in R. Genetic diversity statistics including inbreeding coefficient (FIS) and fixation index (FST) were analyzed using *dartR* package.

3.2.4.2 Linkage Disequilibrium and haplotype analysis

Linkage disequilibrium analysis was carried out using Tassel version 5.2.93. Results from LD analysis were loaded into R where genome-wide LD decay plot was developed using base R functions. The *Haploview* software version 4.1 was used to implement the haplotype analysis (Barrett et al., 2005). Marker data was converted into the linkage format, which is a Plink data format using *blupADC* package in R. The resulting pedigree and map files were converted to .ped and .info formats respectively using Notepad++ software. Analysis was carried out using Haploview default setting where pairwise comparison of markers >500 kb were ignored and individuals with >50% missing genotypes excluded.

Three different approaches to building haplotype blocks were considered: (i) the Solid Spine of LD method (Barrett et al., 2005); (ii) the Four-gamete Rule method (Wang et al., 2002); and (iii) the confidence intervals of the D´method (Gabriel et al., 2002). These techniques were chosen because they have unique features and are frequently used in the construction of haplotype blocks. They rely on gamete frequency and linkage disequilibrium (LD) aiming to create significant blocks within populations and model historical recombination hotspots.

a. Confidence Intervals

This approach, which comes from Gabriel et al. (2002), is the default one in the *Haploview* package. It generates 95% confidence bounds on D prime and labels each comparison as "strong LD," "inconclusive," or "strong recombination." If 95% of informative comparisons are "strong LD" (i.e., not inconclusive), a block is formed. By default, this approach disregards markers with MAF < 0.05. It is possible to modify the MAF cutoff and the confidence bound cutoffs by selecting "Customize Block Definitions" from the Analysis menu. This definition permits the validity of numerous overlapping blocks. The list of all possible blocks is sorted by default, beginning with the largest, and new blocks are added as long as they don't overlap with blocks that have previously been declared.

b. Four Gamete Rule

According to Wang et al. (2002), the Four Gamete Rule (GAM) divides successive markers into haplotype blocks in the event that no evidence of a previous recombination event is detected between any of the marker pairs in a block. The presence of all four haplotypes of the new marker and any other prior marker with a frequency of at least 1% indicates a historical recombination. The process begins with a new block if this is the case, and a block border is made between those markers.

c. Solid Spine of LD

By calculating the LD between every intrachromosomal marker pair, the Solid Spine of LD approach (SPI), which was first presented by the creators of "Haploview" (Barrett et al., 2005), looks for a spine of strong LD. According to this method, if the pairwise D′ is more than 0.8, two markers on the same chromosome form a block border. The block is made up of all the markers in that window. This makes it possible for intermediate markers to be absent from LD.

3.3 Results

3.3.1 SNP Markers Quality and Distribution

A total of 54,877 SNPs were filtered and yielded 37,419 SNPs each with MAF> 0.05 and a maximum missing SNPs of 0.05. The SNPs were well distributed across the 18 cassava chromosomes (Figure 3.1). Chromosome 01 was the largest stretching up to about 43Mb while chromosome 14 was the shortest stretching up to about 29Mb. SNPs were more dense on the telomeres while the centromere was characterized with less dense SNPs.

Figure 3. 1: SNP density plot showing distribution of SNPs acrosscross 18 chromosomes of cassava within a window size of 1Mb

The frequency of transition mutation was higher (68.59%) than transversion mutation (31.41%) where A/G mutation was the most frequent while G/C was the least frequent (Table 3.1). The transition to transversion (Ts/Tv) mutation ratio was 2.18). Chromosome 1 had the highest number (3076) of SNPs while chromosome 7 had the least number (1517) of SNPs (Table 3.3).

Table 3. 1: Number of SNPs due to point mutation types

Mutation Type	SNP mutation	Number of SNPs	Total SNPs per Mutation type
Transitions	A/G	7487	25,667 (68.59%)
	C/T	5586	
	G/A	5657	
	T/C	6935	
Transversions	A/T	1812	11,752 (31.41%)
	A/C	1393	
	C/G	1180	
	G/T	1492	
	T/G	1464	
	C/A	1630	
	T/A	1744	
	G/C	1039	
	Total	37,419	37419
Transition/Transversion ratio			2.18

3.3.2 Genetic Diversity

Genetic diversity statistics (Table 3.2) revealed that the genetic diversity (GD) or expected heterozygosity (H_e) had a mean of 0.29 and ranged between 0.10 and 0.50 while the observed heterozygosity (H_o) had a mean of 0.30 with a range of 0.21 to 0.38. The average polymorphic information content (PIC) was 0.24 and ranged between 0.10 and 0.38. The minor allele frequency (MAF) averaged to 0.20 and ranged from 0.05 to 0.50. The average coefficient of inbreeding (F_{IS}) was -0.02 that ranged between -0.28 and 0.28.

Table 3. 2: Whole population diversity statistics

	Mean	**Minimum**	**Maximum**
GD /H_e	0.29	0.10	0.50
PIC	0.24	0.10	0.38
MAF	0.20	0.05	0.50
H_o	0.30	0.21	0.38
F_{IS}	-0.02	-0.28	0.28

Genetic diversity statistics averaged per chromosome showed that MAF ranged between 0.19 – 0.22, H_e and H_o ranged between 0.28 – 0.31 and 0.29 – 0.31, respectively, PIC ranged between 0.23-0.25 while *p* and *q* alleles ranged between 0.38 – 0.41 and 0.57- 0.61 respectively (Table 3.3). Majority of the SNPs had PIC > 0.20 and MAF > 0.10 (Figure 3.2)

Table 3. 3: Diversity statistics and SNP distribution in 18 chromosomes of cassava

Chrom.	SNP number	p	q	MAF	He	Ho	GD	PIC
Chr01	3076	0.40	0.60	0.20	0.29	0.30	0.29	0.24
Chr02	2352	0.40	0.60	0.21	0.30	0.30	0.30	0.25
Chr03	2305	0.39	0.61	0.20	0.29	0.29	0.29	0.24
Chr04	2436	0.42	0.58	0.20	0.29	0.29	0.29	0.24
Chr05	2196	0.38	0.62	0.20	0.29	0.30	0.29	0.24
Chr06	2253	0.40	0.60	0.21	0.30	0.31	0.30	0.25
Chr07	1517	0.40	0.60	0.21	0.30	0.31	0.30	0.25
Chr08	1936	0.40	0.60	0.22	0.31	0.31	0.31	0.25
Chr09	2069	0.40	0.60	0.19	0.28	0.29	0.28	0.24
Chr10	2015	0.42	0.58	0.21	0.30	0.30	0.30	0.25
Chr11	2209	0.41	0.59	0.21	0.30	0.30	0.30	0.25
Chr12	1746	0.41	0.59	0.21	0.30	0.31	0.30	0.25
Chr13	1973	0.38	0.62	0.19	0.28	0.29	0.28	0.23
Chr14	2236	0.41	0.59	0.20	0.29	0.30	0.29	0.24
Chr15	2087	0.39	0.61	0.20	0.30	0.31	0.30	0.24
Chr16	1554	0.38	0.62	0.21	0.30	0.31	0.30	0.25
Chr17	1873	0.43	0.57	0.21	0.30	0.30	0.30	0.25
Chr18	1586	0.39	0.61	0.20	0.29	0.30	0.29	0.24

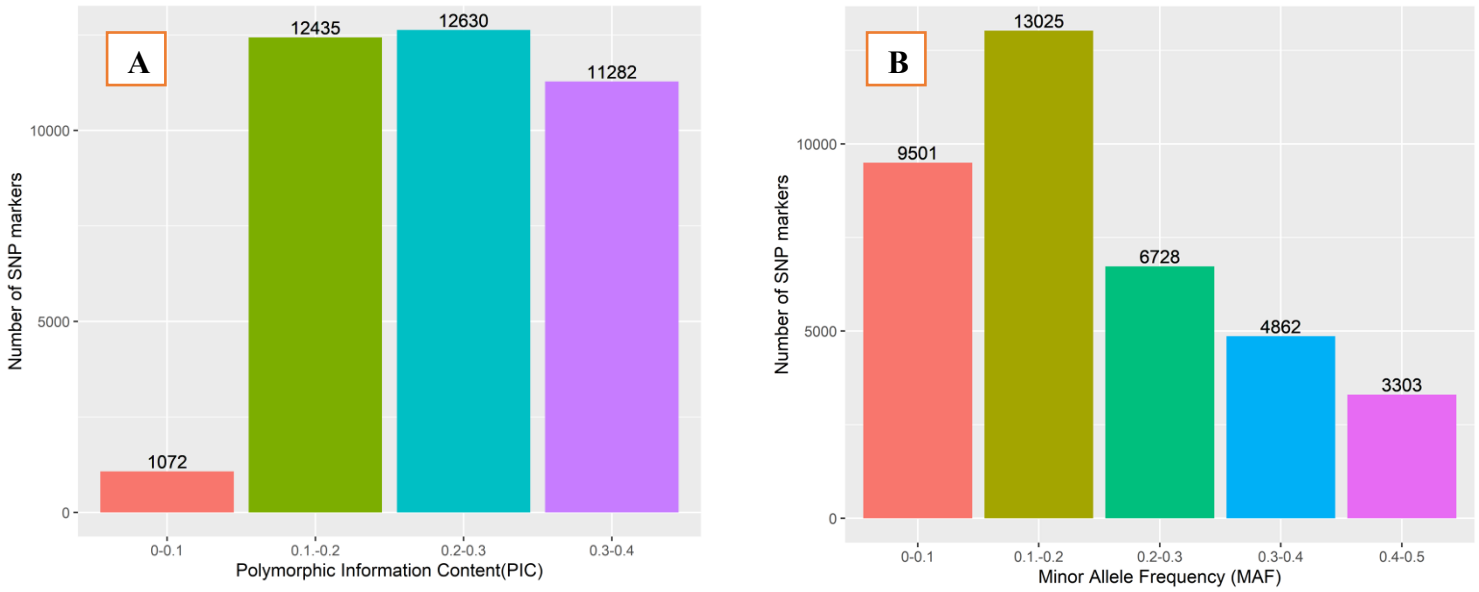


Figure 3. 3: Plot of SNP marker quality analyses showing number of markers with A (polymorphic information content of 0.0-0.1, 0.1-0.2, 0.2-0.3, 0.3-0.4) and B (minor allele frequency of 0.05-0.1, 0.1-0.2, 0.2-0.3, 0.3-0.4, 0.4-0.5)

3.3.3. Population Structure

Population structure analysis indicated that out of possible 1-10 ancestral populations (K), the most significant change was observed at K=3. This result showed that the 93 pVAC genotypes can optimally be classified into three ancestral populations. The results showed that 32, 28 and 33 genotypes were assigned into cluster 1, 2 and 3 respectively. Hierarchical dendrogram visualized the result confirming the three ancestral populations (Figure 3.3).

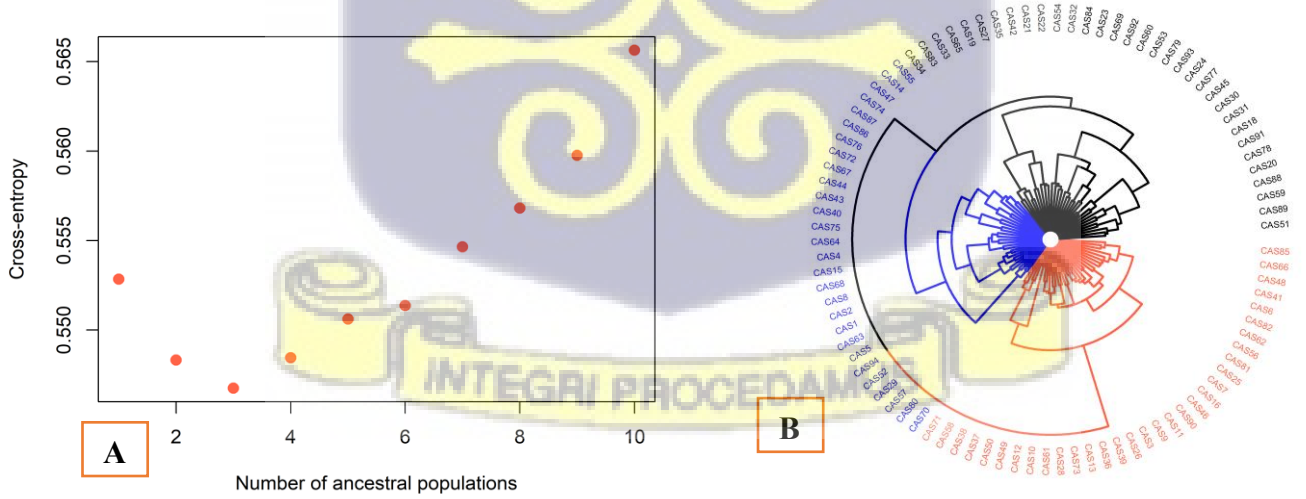


Figure 3. 5: Plot of the optimum number of ancestral populations (K) from cross-entropy criterion showing K=3 (A) and hierarchical dendrogram visualizing the clusters (B)

Within and among population variability was determined using analysis of molecular variance (AMOVA). This analysis revealed that 96.05% of genetic variability existed within populations while 3.95% existed between populations (Table 3.4). This result points out that genetic differentiation within subpopulations was higher than between subpopulations.

Table 3. 4: Analysis of Molecular Variance (AMOVA) of the three ancestral populations

	Df	**Sum of Square**	**Mean Square**	**Variation (%)**
Between populations	2	71696.71	35848.36	3.95
Within populations	90	1420082.06	15778.69	96.05
Total	92	1491778.77	16214.99**	100.00

The intercluster fixation index (Fst) were about zero in all cluster combinations (Table 3.5). This reveals that the populations are interbreeding freely. Individuals in the three ancestral populations exhibit considerable heterogeneity in estimated admixture proportions (Figure 3.4).

Table 3. 5: Fixation index (F_{st}) among the three subpopulations

	Subpopulation 1	Subpopulation 2	Subpopulation 3
Subpopulation 1	0		
Subpopulation 2	0.000922	0	
Subpopulation 3	0.001052	-3.54E-05	0

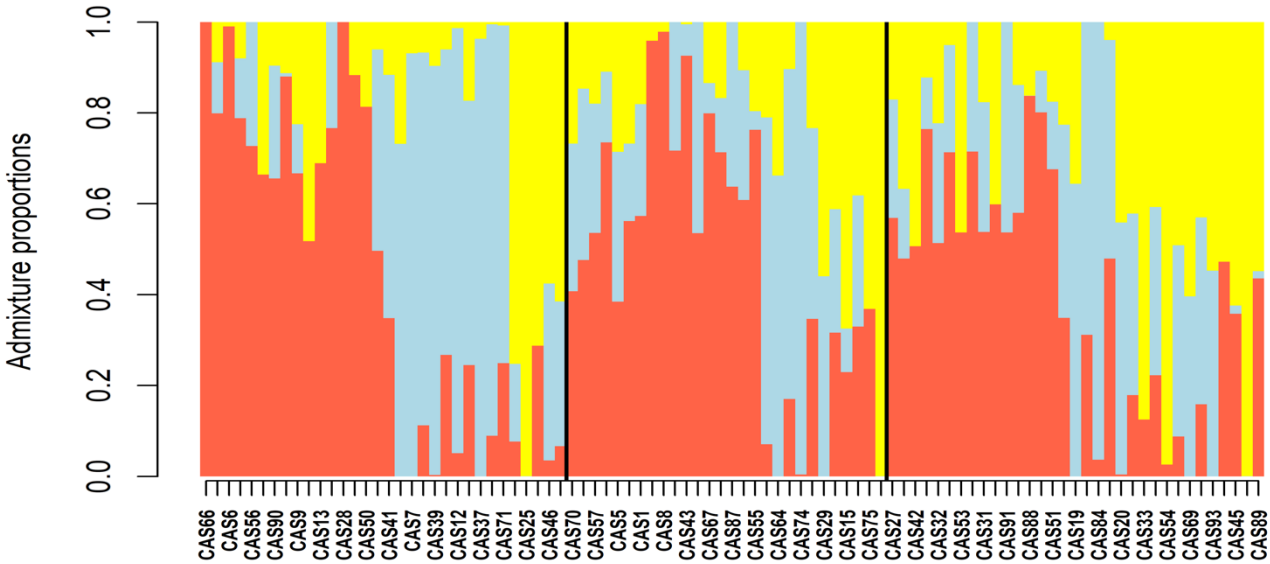


Figure 3. 6: Admixture proportion of genotypes in the three ancestral populations

Out of 37,419 loci, 34,085, 33,784 and 34,324 polymorphic loci were observed in population 1, 2 and 3 respectively. The three sub populations had equal H_o and H_e whereby H_o was higher than H_e. The inbreeding coefficient (F_{IS}) appeared to be similar ranging from -0.38 to -0.39 (Table 3.6).

Table 3. 6: Diversity statistics of the three ancestral subpopulations

pop	nInd	nLoc	polyLoc	monoLoc	H_o	H_e	F_{IS}
1	32	37419	34085	3334	0.36	0.26	-0.38
2	28	37419	33784	3635	0.36	0.26	-0.38
3	33	37419	34324	3095	0.36	0.26	-0.39

Pop= population, nInd = number of individuals, nLoc = number of loci, polyLoc = polymorphic loci, monoLoc = monomorphic loci, H_o = observed heterozygosity, H_e= expected heterozygosity, F_{IS} = coefficient of inbreeding

3.3.4. Linkage Disequilibrium

The genome-wide LD (indicated by r^2) decreased with physical distance between SNPs. The LD decay level measured is the chromosomal distance when LD decreased to 0.2 or 0.1. The results showed that at a threshold of r^2=0.2, LD decayed at 613.072 kb while at r^2=0.1, LD decayed at 1786.714kb (Figure 3.5).

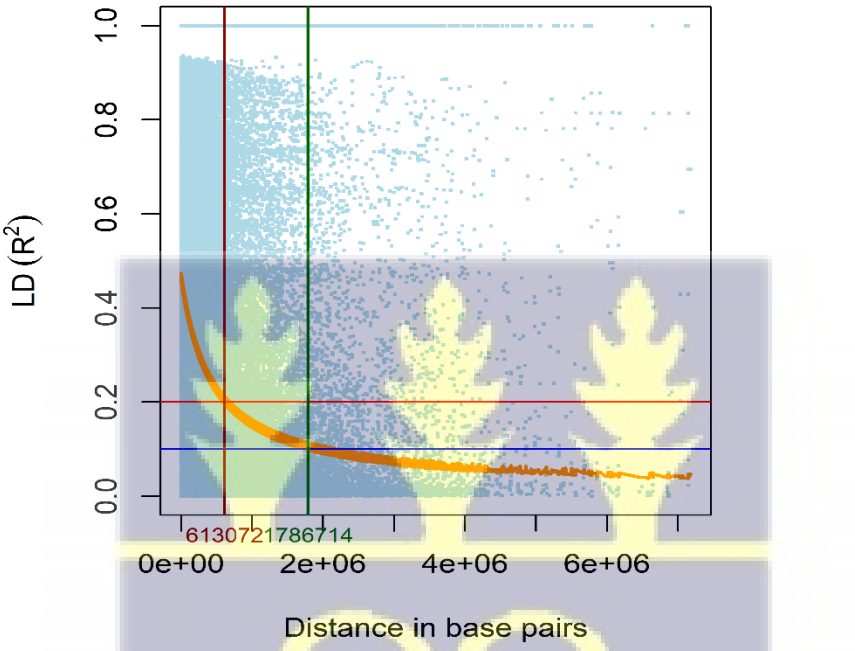


Figure 3. 7: Genome-wise Linkage disequilibrium (LD) decay plot

3.3.4. Haplotype

Within all datasets using the methods implemented in *Haploview*, the number of haplotype blocks was consistently lower than the number of total SNPs. The haplotype methods yielded different results with Confident Intervals method yielding the lowest number while Four Gamete Rule method yielding the highest number of haplotypes across the cassava genome.

Table 3. 7: Number of haplotype blocks identified across the 19 chromosomes of cassava genome using Confidence Intervals, Four gamete rule and Solid Spine of LD methods

Chrom.	SNP number	Confidence Intervals	Four Gamete Rule	Solid Spine of LD
1	3076	100	392	333
2	2352	83	265	200
3	2305	93	248	151
4	2436	76	244	202
5	2196	91	233	165
6	2253	79	259	178
7	1517	54	174	131
8	1936	63	219	163
9	2069	91	226	161
10	2015	65	222	172
11	2209	80	235	170
12	1746	65	211	140
13	1973	68	216	166
14	2236	85	246	178
15	2087	88	238	161
16	1554	59	167	59
17	1873	74	222	170
18	1586	48	178	131
Total	37,419	1,362	4,195	3031

Chrom. = Chromosome number

3.4. Discussion

The current study is the first to investigate population structure, genetic diversity, extent of LD and haplotypes present in pVAC cassava population using DArTseq SNP markers in Kenya. The SNP markers were well distributed on all 18 chromosomes of the cassava genome. SNPs were denser at the telomere while in the centromere they were less dense or absent. This is attributed to the fact that recombination is expected to be higher at the telomere while at the centromere it is expected to be suppressed (Carneiro et al., 2009). Reports indicate that loci near centromere exhibits a high level of linkage disequilibrium than those in the telomere, facilitating high recombination on telomeres compared to centromeres (Carneiro et al., 2009). It has been reported that marker density is one of the factors that affect the accuracy of genomic studies (Desta & Ortiz,

2014). It has been proposed that high-density markers can increase genomic predictive accuracy, and it is generally accepted that more markers lead to higher accuracy up to a plateau (Wenjie et al., 2024).

Results showed that 68.59% of SNPs were due to transition mutation (purine-to-purine or pyrimidine-to-pyrimidine mutation) while 31.41% being due to transversion mutation (purine-to-pyrimidine mutation or vice versa) with transitition/transversion ratio of 2.18. This result is similar to other reports documenting that transition mutations are more frequent than transversion mutations (Aloqalaa et al., 2019). Reports indicate that selection does not favour transversions because they are unlikely to keep biochemical properties of the initial amino acid or protein (Lyons & Lauring, 2017). The structure of the genetic code makes nonsynonymous transitions unlikely to cause radical change on protein like transversion. The results are in line with reports that naturally, transition mutations are more frequent than transversions in populations (Aloqalaa et al., 2019).

The study population had a transition to transversion (Ts/Tv) ratio of 2.18. One of the most important metrics for researching genetic diversity is the transition/transversion (Ts/Tv) ratio. It offers information about evolutionary history both within and across populations, selection pressures, and mutation processes (Moore & DeFillippis, 1997). A lower Ts/Tv ratio suggests increased transversions, which may be a sign that selection is favoring beneficial nonsynonymous mutations. In protein-coding genes, adaptive mutations often require amino acid changes, which are more likely to occur through transversions. Therefore, transversion mutation leads to new alleles, which contributes to genetic diversity in a population. Based on this, a lower Ts/Tv ratio suggests that the population has mutation diversity which contributes to increased genetic diversity. Other studies working on highly diverse cassava populations have reported a lower

Ts/Tv ratio (1.06 - 1.54) compared to 2.18 in the current study population (Neimsemman et al., 2024; Sesay et al., 2023; Soto et al., 2015). This result suggests that mutation diversity in the study population was moderate.

The population's marker polymorphic information content (PIC) ranged between 0.10 and 0.38 with a mean of 0.24. Markers with PIC values of less than 0.20 are less informative and this can be attributed to the bi-allelic nature of SNP marker which limits them to a maximum PIC value of 0.5 (Z. Luo et al., 2020). However, the study population has a mean PIC of 0.24 with majority of the markers ranging between 0.2 and 0.4 and this made the markers reliable for use in this study (Serrote et al., 2020).

The MAF ranged between 0.05 – 0.50 with majority having MAF > 0.10. Minor alleles refer to the less common (rare) alleles of SNP in the genome. In genomic studies, markers with low MAF are removed so as to increase the frequency of alleles that have important effect of the trait of interest. Anderson et al. (2010) reported that failing to remove low-frequency alleles from genomic data leads to results with false information for two reasons: 1) the associations observed in these SNPs are small because they are addressed by the genotypes of a small number of individuals; and 2) they may be caused by genotyping errors in markers that are actually monomorphic in the population. It is with this regard that MAF < 0.05 (5%) are often removed to increase the genotype quality during the development of genomic models. With majority of the markers having MAF > 0.10 (10%), the genotype quality in the present data was reliable.

Average H_e is the probability that at a single locus of a diploid organism, any two alleles chosen at random are different (Mukhopadhyay & Bhattacharjee, 2016; Serrote et al., 2020). Unlike H_o, H_e is less sensitive to population size and thus, preferred in determining population genetic diversity (Mukhopadhyay & Bhattacharjee, 2016). The GD or H_e in the whole pVAC cassava

population was 0.29 and this was lower to H_o (0.30). These results show a lower H_e value compared to H_e=0.5 reported by Esuma et al. (2012) who studied population diversity of Ugandan pVAC population using SSR markers. Cassava is a highly heterozygous plant and therefore, an average H_e of 0.29 signifies relatively low diversity. This is attributable to the fact that the study population was composed of progenies from parents that were partial inbred lines (S_2) that had gone through two cycles of selection (C2). Nevertheless, H_e ranged between 0.10 and 0.50 while H_o ranged from 0.21-0.38 which was similar to the results reported by Kamanda et al. (2020). This implies that the population has diverse alleles and thus, it can be used for breeding cassava with high pVAC. A lower H_e (0.29) compared to H_o (0.30) in the study population implies that there were more heterozygotes in the study population than expected in a population undergoing a random mating (Mukhopadhyay & Bhattacharjee, 2016). Therefore, these results signify an absence of random mating in the pVAC cassava population. The average F_{IS} of the whole population was -0.02, implying that there were slightly higher heterozygotes. This slightly lower than zero F_{IS} could be attributed to the fact that the population under study was composed of half-sib progeny from partially inbred parents. A F_{IS} value of -0.02 may be a small effect but suggests the absence of random mating in the study population. Genetic diversity statistics did not differ among the 18 chromosomes of cassava genome. Each chromosome had nearly similar diversity statistics values that were similar to the population's diversity.

The study population was structured into three subpopulations with 32, 28 and 33 genotypes respectively for subpopulations 1, 2 and 3. Genetic variability among and within sub populations was determined by AMOVA where variability within subpopulation was higher (96.05%) than that of between subpopulations (3.95 %). A number of studies have documented a higher variability within population than between populations in different crops (Meira et al., 2019;

Souza et al., 2019). According to Souza et al. (2019) higher within population variability is due to sampling of more genotypes per population during germplasm collection. This high variation within populations can serve as a source of superior population selection (Meira et al., 2019). Other reports indicate that woody and outcrossing plants generally harbor high variation within populations (Sheng et al., 2005). High within population and low between population variability is due to high gene flow due to outcrossing nature of crops such as cassava (Sheng et al., 2005).

Significantly greater genetic variability within populations than between populations, as indicated by AMOVA results, suggests that most genetic diversity is found within individual populations rather than between them. This suggests a high level of gene flow and possibly recent population expansion, with little genetic differentiation between groups (Sheng et al., 2005). A high degree of genetic exchange between populations prevents different genetic clusters from emerging, resulting in a homogeneous gene pool across the species range. Since the majority of genetic diversity is located within each population than between population, conservation efforts should prioritize the maintenance of local genetic diversity rather than focused on maintaining distinctions across populations.

Genetic differentiation of the subpopulation was determined by inter-cluster fixation index (F_{st}) which measures genetic differences across populations (Smaragdov & Kudinov, 2020). The F_{st} values range from 0 to 1 where 0 means that there is complete sharing of alleles (panmixia) while 1 means that there is no sharing (fixed). The three subpopulations in the study population had inter-cluster F_{st} values of near zero, and this ranged from -3.54E-05 to 0.001052. This revealed that the three subpopulations shared alleles to a large extent and this is supported by AMOVA results where variability between populations was very low.

Intra-cluster diversity analysis results indicates that all the subpopulations had equal H_e (0.26) and equal H_o (0.36). However, H_o was larger than H_e in the three subpopulations, indicating the presence of high heterozygosity than expected in a random mating population. These results suggests that the subpopulations are in non-random mating but in a way to avoid inbreeding (Mukhopadhyay & Bhattacharjee, 2016). Furthermore, the coefficient of inbreeding (F_{IS}) across the subpopulations were -0.38, -0.39 and -0.39 for subpopulation 1, 2 and 3 respectively. F_{IS} is the average SNP homozygosity within an individual relative to expected allele homozygosity. F_{IS} values range from -1 to 1 where 1 indicates high degree of inbreeding and deficit of heterozygotes while -1 indicate excess heterozygotes and thus no inbreeding. The subpopulations in the current study had F_{IS} values approaching -1 and therefore, the three subpopulations had excess of heterozygotes. This is expected since cassava is a cross-pollinated crop undergoing outbreeding and thus highly heterozygous.

LD analysis is an important statistics in genomic studies whereby LD decay reveals the population recombination history and helps in determining the number of markers required to achieve a high resolution of genomic studies (Zhang et al., 2019). Results revealed a low mean LD among markers that decayed slowly over a long physical distance. The slow LD decay in cassava is unlikely; however, this was possible because the genotypes used in the present study were from progenitors that were partial inbred lines. Results showed a genome-wide LD decay at a physical distance of 613. 072 kb at a threshold of $r^2 = 0.2$. This result shows that there is a high possibility of carrying out association mapping and putative gene identification in the study population. Furthermore, LD decay provided information on the number of markers that can reliably be used in genomic studies. Results from this study revealed an LD decay at 613.072kb (0.613Mb) which means that for high resolution genomic analysis on the population, there should be at least one marker for every

0.613Mb. Considering that the size of cassava genome is estimated to extend to 770Mb, an LD decay of 0.613Mb means that about 1,256 markers that are well distributed can fully cover the whole genome. This information is helpful in making decision on the depth of genotyping required for genomic studies.

Results show that different methods for haplotype analysis identified varying numbers of haplotype blocks. These varying results among haplotype identification algorithms can be exploited to achieve optimum results from genomic analysis. It was noted that the number of haplotype blocks were few compared to the number of SNP markers across all 18 chromosomes of cassava genome. This is of importance in genomic analysis using machine learning algorithms that require reduction of data complexity. Data complexity can be decreased by combining thousands of SNPs into a few hundred haplotype blocks. Genomic analyses involving fewer haplotype blocks lowers the number of statistical tests and the possibility of incorrect associations between trait and markers. Cassava breeders can utilize this data as a foundation for haplotype-based genomic research. The use of haplotype-based genomic selection has been shown to improve the prediction accuracy up to 7% in predicting soybean yield (Yoosefzadeh-Najafabadi et al., 2022).

3.5. Conclusion

The present study aimed at determining genetic diversity, population structure, extent of LD decay and haplotypes in a pVAC cassava population in Kenya. The marker data used in this study was informative and well distributed across the cassava genome. Transition mutations were more frequent than transversion mutations in the study population. Genetic diversity results revealed presence of modest genetic diversity in the study population. There were moderately diverse alleles

in the population that can be used for development of superior pVAC genotypes. Mating in the pVAC cassava population was more probably random without inbreeding.

The study population was structured into three subpopulations where variability within a subpopulation was higher compared to between subpopulations. The subpopulations were in total panmixia and thus, they were interbreeding freely. Each of the three subpopulations had higher observed heterozygosity compared to expected heterozygosity. This indicates the presence of genetic variability that cassava breeders can exploit to achieve heterosis.

The study population had a low mean LD among markers that decayed slowly over a long physical distance (1786.714kb) at a threshold of r^2= 0.2. The number of haplotype blocks in the study population was consistently lower than the number of total SNPs. The three haplotype analysis methods yielded different results where the Confident Intervals method resulted to the lowest number (1,362) while Four Gamete Rule method resulted in the highest number (4,195) of haplotypes across the cassava genome

CHAPTER FOUR

4.0. GENOMIC PREDICTION MODELS FOR PRO-VITAMIN A CASSAVA BREEDING IN KENYA

4.1. Introduction

Global efforts to biofortify cassava with pVAC are crucial in addressing vitamin A deficiency (VAD) among individuals who depend on the crop as their staple. One of the key challenges that face these biofortification efforts is slow growth rate of cassava, where the crop matures at about 12 months (Augusto & Alves, 2002). This long maturity time for the crop is a constraint to plant breeders because it takes a long time to evaluate and select parental genotypes with promising root traits (Ceballos et al., 2020). The challenge has created the need for a system that reduces selection cycle time through selection of progenitors at early stages. Genomic prediction (GP) also known as genomic selection (GS) has been suggested for use in shortening cassava selection cycle time (Voss-Fels et al., 2019).

GP uses genomic models to predict individual's genomic estimated breeding value (GEBV). These models are developed by associating molecular markers with phenotypic information. The GEBVs are additive genetic effects of individuals that are used to select promising parents in a breeding population. GP accelerate genetic gains by shortening selection cycle time and increasing selection accuracy and intensity (Jannink et al., 2010). Accelerated genetic gains leads to quick improvement of crop varieties and fast response to the need of farmers. Development of GP models requires the use of high-density genetic markers to ensure that all causal genes are in LD with at least one marker. Nevertheless, information on the extent of LD decay in the population, guide on the minimum marker density to be used in genomic analysis (Karimi et al., 2020; Tsetsos et al., 2018).

Performance of genomic selection in plant breeding is affected by trait heritability, model accuracy, relatedness between training and validation sets of genetic resources, genetic architecture of the trait, marker density and extent of LD between markers (Desta & Ortiz, 2014). These factors determine the accuracy of GEBV and thus, affect the response of genomic selection in plant breeding. The genetic background of traits of interest may present complex genetic architecture that affect the accuracy of GP models (Mota et al., 2024).

Most GP models are parametric and thus developed from algorithms assuming that the target trait is influenced by additive genetic effects from multiple loci distributed across a genome (Ren et al., 2021). Traditionally, GP exploit markers by parametric models such as genomic best linear unbiased prediction (GBLUP), and Bayesian regressions (Howard et al., 2014). GBLUP approach assume that the observed trait phenotype is influenced by several loci exhibiting additive genetic effects across the genome (Ren et al., 2021). Bayesian regression approaches assign different weights to markers so as to align with the trait's genetic architecture (Pérez & de los Campos, 2014). These parametric GP approaches assume a linear association of marker effects and phenotype and thus ignores possibility of non-additive genetic effects (dominance and epistasis) (Momen et al., 2018). Despite this, it has been reported that traits may be controlled by complex genetic architecture involving additive and non-additive genetic effects (Varona et al., 2018). Therefore, statistical approaches that account for additive and non-additive genetic effects to predict the total genetic value of individuals are needed.

Recently, research directed toward modelling traits to accommodate additive and non-additive genetic effects has been given great attention (Mota et al., 2024). It has been suggested that these statistical methods could improve GP accuracy. Reports indicate that the use of machine learning (ML) models provide the capacity to reliably accommodate complex association between trait

phenotype and predictor features, accounting for both additive and non-additive effects (Y. Liu et al., 2019).

The parametric GP model is expressed as $y = X\beta+Z\mu+\varepsilon$, where β and μ represent fixed and random effects, respectively. Adjusting the fixed and random factors in GP models could improve their performance. For example, Li et al. (2019) demonstrated that modeling major-effect SNPs as fixed effects improved genome prediction. In wheat stem rust resistance studies, it was discovered that a GBLUP model including markers linked with stem rust resistance as fixed effects outperformed a basic GBLUP model (Rutkoski et al., 2014).

Research on the effectiveness of GP in biofortification of cassava with pVAC revealed that incorporating chromosome 1 markers that are associated with the trait improved accuracy by r = 0.06 (from r = 0.52 to r = 0.58) (Esuma et al., 2021). However, the tested models were parametric models that only accommodate additive genetic effects. Moreover, the incorporated SNPs were identified using traditional single locus GWAS which has lower statistical power compared to multi-locus random marker effect GWAS (mr-GWAS) (Wang et al., 2016). Because of its high statistical power, mr-GWAS implementation can, in fact, result in the discovery of additional genomic signatures from marker-trait association in small populations. The multi-locus random-SNP-effect mixed linear model (mrMLM), which treats the SNP-effect as random, is used to implement the mr-GWAS (S. B. Wang et al., 2016; Y. W. Zhang et al., 2020).

The estimated marker effects are reduced towards zero under the random GWAS model. In this model, a modified Bonferroni correction is used to estimate the threshold p-value for genome-wide significance testing. According to Wang et al. (2016) the modified Bonferroni correction threshold p-value is 0.05/Me, where Me is the effective number of markers. This method uses an effective amount of markers to account for multiple testing. Currently, there is no report on the

incorporation of SNPs from mr-GWAS into parameterized GP model for improved accuracy. Similarly, studies on the use of ML algorithms for prediction of total genetic value of pVAC cassava genetic resource are yet to be reported.

The objectives of this study were to:

a) determine the strategies for improving the prediction performance of parametric and non-parametric GP models for prediction of pVAC in cassava roots; and
b) determine the effectiveness of modeling additive and non-additive genetic effects in pVAC cassava using non-parametric ML models.

4.2. Materials and Methods

4.2.1. Genetic Resources

A panel of 94 pVAC cassava genotypes were used in this study. These genotypes are part of 120 accessions from National Crop Resources Research Institute (NaCRRI), Uganda. The genetic resources were composed of half-sib progenies of parents from cycle 2 selection of pVAC cassava partial inbred lines. The accessions (in form of seeds) were germinated in a screen house and the resultant seedlings transplanted as training population at KALRO-NRI, Kakamega, Kenya. Since pVAC is a high heritability trait (h^2=0.72) and selection takes place at seedling stage under rapid cycling recurrent selection scheme (Ceballos et al., 2013), the seedling population was used as the training population. From this population, the 94 genotypes (2 white rooted and 92 yellow rooted) were selected based on root flesh colour using qualitative colour chart in a scale of 1-8 (Figure 4.1).

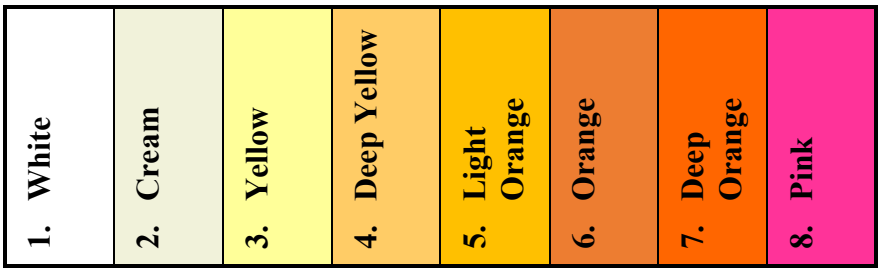


Figure 4. 1: Qualitative colour chart with assigned values on a scale of 1-8 for carotenoid content

4.2.2. Phenotyping

Two sample roots from each genotype were chosen at harvest (12 months after planting). The roots were wrapped in aluminum foil, labeled, and promptly sent to the University of Nairobi's Department of Food Science, Nutrition, and Technology Laboratory in Nairobi, Kenya for beta-carotene analysis. Given the degrading nature of carotenoids, harvesting was done in the evening under low light, and the sample was delivered to the laboratory at night, within 8 hours of harvest. The roots were cleaned with tap water, wiped, then peeled. Each peeled root was cut twice longitudinally with a kitchen knife, yielding four pieces. Two opposing pieces were chosen from each root, yielding a total of four pieces from two roots per genotype. The chosen pieces were further diced into tiny pieces, properly combined, and two portions, each weighing 50g of the mixture was selected as a representative sample for analysis. Using the two replications of a sample (two portions of sample weighing 50g each) beta-carotene content in the 94 cassava root samples were determined using high performance liquid chromatography as described by Rodriguez-amaya & Kimura (2004).

4.2.3. Genotyping and SNP quality control

Genotyping of the genetic resources was carried out using DArTseq (GBS) (1.0) genotyping as described in section 3.2.2. SNP quality control was achieved by filtering out MAF < 0.05 and a

SNP call rate of 0.95 where SNPs with more than 5% missing data were removed using *snpReady* package in R statistical software.

4.2.4. Identification of significant SNP marker effects

Association studies were carried out to identify SNPs with significantly higher effects for use in optimizing genomic selection models. Genome wide association analysis was carried out using two different multi-locus GWAS algorithms. First, multi-locus fixed marker effects GWAS with its conservative and stringent Bonferroni correction was implemented using BLINK model in GAPIT v3.1.0 package (Wang & Zhang, 2021) in R. Secondly, a multi-locus random marker effects GWAS, which adopts a modified threshold p-values, was implemented using four models (mrMLM, FASTmrMLM, pLARmEB and ISIS EM-BLASSO) in mrMLM v4.0.2 package (Zhou et al., 2021) in R.

4.2.5. Genomic Prediction Models

Parametric Bayesian generalized linear (BGLR) GP models that included Bayesian Ridge Regression (BRR), Bayesian Lasso (BL), BayesA, BayesB, and BayesC were developed using *BGLR* package in R. These models can be expressed as:

$$\boldsymbol{y = 1\mu + \chi_1\beta_1 + \varepsilon}$$

Where; $\mathbf{y}$ = phenotype vector of beta-carotene content, $\mathbf{1}$ = Incidence matrix for fixed effects (intercept), $\boldsymbol{\mu}$ = Vector of fixed effects coefficients (GWAS significant SNPs), $\boldsymbol{\chi_1}$ = Incidence matrix of random effect, $\boldsymbol{\beta_1}$ = Vector of random SNP effects coefficients: $\boldsymbol{\beta_1} \sim \mathbf{N}\ (\mathbf{0}, \mathbf{I}\sigma^2_{\beta 1})$, $\boldsymbol{\varepsilon}$ = Vector of residual terms: $\boldsymbol{\varepsilon} \sim \mathbf{N}\ (\mathbf{0}, \mathbf{I}\sigma^2_{\varepsilon})$.

To enhance the prediction performance of the developed parametric GP models, GWAS significant SNPs from multi-locus fixed marker effect and multi-locus random marker effect GWAS were

fitted into the BGLR model as fixed effects. The GWAS significant SNPs were fitted as a covariate in the BGLR model above, where μ is the vector of these SNP effects.

Non-parametric GP models that accommodate additive and non-additive genetic effects were fitted using ML models in *Caret* package in R. These models included Random Forest (RF), Support Vector Machine (SVM), Neural Networks (NNET), Extreme Gradient Boosting (XGBOOST), K-Nearest Neighbors (KNN) and the ensemble of these models. To remove redundant predictors, feature/variable selection was carried out using *Boruta* package in R. These models can be expressed as:

$$\mathbf{y=X\beta+K\alpha+\varepsilon}$$

Where; **y** = Phenotype vector of beta-carotene content, **X** = Genomic marker matrix, **β** = Vector of linear coefficients of additive marker effects, **K** = kernel matrix of non-linear relationships, **α** = Vector of non-linear coefficients of non-additive marker effects: $\boldsymbol{\varepsilon \sim N\ (0,\ 1\sigma_{\varepsilon}^{2})}$.

4.2.6. Model evaluation

The K fold cross-validation procedure was employed to assess the models' performance (Nti et al., 2021). Five folds, or parts, comprising 20% of the whole dataset, were randomly selected from the data. Every fold was utilized as a testing set one at a time, and the remaining four sets, or 80% of the data, were used to train the model. The 5-fold cross-validation was repeated 100 times to generate robust predictions. The prediction ability of each model was evaluated using the correlation coefficient (r) between predicted (GEBV of genomic values) and actual phenotypic values. The percentage improvement of the model was computed using the following formula:

$$Improvement\ (\%) = \frac{r_2 - r_1}{r_1} \times 100$$

Where, r_2 = prediction ability after model improvement, r_1 = prediction ability before model improvement

4.3 Results

4.3.1 Phenotypic information

The pVAC training population had all trans Beta-carotene (ATB-Carotene) content with a mean of 5.20 µg/g, a minimum value of 0.40 µg/g and a maximum value of 10.30 µg/g (Table 4.1). Based on the colour scale values, the population had a minimum value of 1, a maximum value of 6 and a mean of 3.62 (approximately 4).

Table 4. 1: Beta-carotene content and root colour of pVAC training population

n = 94	Mean	SD	Median	Min	Max	CV
B-Carotene (µg/g)	5.20	2.40	4.70	0.40	10.30	46.37
Root colour (Score values)	3.62	1.03	4.00	1.00	6.00	28.33

SD = Standard Deviation, Min = Minimum value, Max = Maximum value, CV =coefficient of variation, n= number of genotypes

Density plot in Figure 4.2 shows that the phenotypes followed a normal distribution. Majority of the samples had beta-carotene content greater than 0.50mg/100g (5.0µg/g) and the beta carotene levels among the genotypes did not fall into one category and thus exhibited some variation. The genotypes had some level of genetic relationship and less structured as shown on genomic relationship heatmap

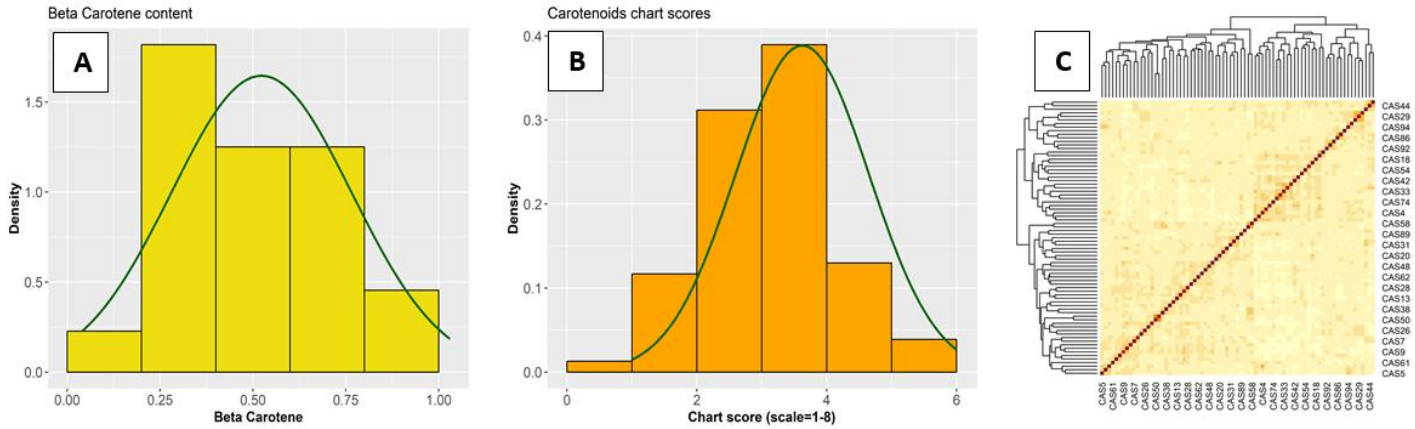


Figure 4. 2: Characteristics of the genetic resources used in this study. A is the density plot of beta-carotene values while B is the density plot of root flesh colour chart scores. C is the heatmap of the genomic relationship matrix

4.3.2 Identification of significant SNP marker effects

4.3.2.1 Multi-locus fixed marker effect model

The multi-locus fixed marker effect GWAS results from BLINK model identified one significant SNP associated with cassava root flesh colour (Table 4.2). This significant SNP was located on chromosome 01 position 31936067. The models detected two significant SNPs associated with cassava root beta-carotene content on chromosomes 09 and 14, respectively, at positions 34394428 and 6077652.

Table 4. 2: Significant SNPs identified by BLINK GWAS model

Carotenoid	GWAS Model	Chromosome	Position bp.
All trans β-carotene	BLINK	09	34394428
All trans β -carotene	BLINK	14	6077652
Root flesh colour	BLINK	01	31936067

bp. = Base pairs

The multi-locus random marker effect GWAS identified five significant SNPs on chromosomes 01, 03, 04, 14 and 18 associated with beta-carotene content in cassava roots (Table 4.3). The results indicate that mrMLM model identified four significant SNPs on chromosomes 01, 04, 14 and 18. Among the four models of multi-locus random marker effect GWAS, it is only mrMLM that was able to detect a significant SNP in chromosome 01. Similarly, ISIS EM-BLASSO model identified 3 significant SNPs on chromosomes 04, 14 and 18. Both FASTmrMLM and pLARmEB models identified two significant SNPs on chromosomes 03 and 14.

Table 4. 3: Significant SNPs resulting from mrMLM GWAS models

Method	Chrom.	Position (bp)
mrMLM	01	5355523
mrMLM & ISIS EM-BLASSO	04	7941731
mrMLM, FASTmrMLM, pLARmEB & ISIS EM-BLASSO	14	6077652
mrMLM & ISIS EM-BLASSO	18	24415388
FASTmrMLM & pLARmEB	03	5342366

Chrom. = Chromosome, bp = Base pairs

4.3.3 Genomic Prediction Models

4.3.3.1 Parametric GP models

Prediction ability of parametric GP model for predicting colour score value improved when the single GWAS significant SNP was added into the model as fixed effects (Table 4.4). Among the basic models, Bayesian LASSO (BL) model had the highest prediction ability (r= 0.48) whereas Bayesian Ridge Regression (BRR) model had the lowest prediction ability (r = 0.39). However, after adding significant SNP as fixed effects, the prediction ability of all models improved whereby, the highest prediction ability was r = 0.63 from BRR and lowest was r = 0.61 from BL,

BayesA and BayesB. This prediction ability improvement resulted in BRR having the highest improvement (62%) and BL having the lowest improvement (27%).

Table 4. 4: Prediction ability of parametric models developed from BGLR models for predicting root flesh colour score values before and after adding significant SNPs as fixed effects in the model, and the percentage improvement of the models after adding significant SNPs. The prediction performance values are correlation coefficients between predicted and actual values

	Model 1		Model 2a		% Improvement
	Training	**Validation**	**Training**	**Validation**	**Validation**
BRR	0.98	0.39	0.91	0.63	62 %
BL	0.96	0.48	0.73	0.61	27%
BayesA	0.99	0.43	0.90	0.61	42%
BayesB	0.98	0.42	0.94	0.61	45%
BayesC	0.98	0.44	0.92	0.62	42%

Model1 = Initial GP model, Model 2a = GP model after single significant SNP from GWAS as fixed effect

Prediction ability of GP model for prediction of beta-carotene in cassava root improved when the two significant SNPs from BLINK model and five significant SNPs from mrMLM models were added as fixed effects to initial the basic BGLR models (Table 4.5). The basic BGLR models had BayesA having the highest prediction ability of r = 0.16 for beta-carotene whereas BRR, BL and BayesB had the least prediction ability of r = 0.14. When the two significant SNPs from BLINK GWAS model were added to the basic BGLR models as fixed effects, the percentage improvement of the models ranged from 163% for BayesA to 207% for BL. Results indicate that after the improvements, BL achieved the highest prediction ability (r = 0.43) while the lowest prediction ability was achieved by BRR and Bayes B, both with r = 0.41. When the five significant SNPs from mrMLM model were added to the basic BGLR models as fixed effect, the percentage improvement of the BGLR models ranged from 350% for BayesA to 479% for BL. Results from

the new model indicated that BL achieved the highest prediction ability (r = 0.81) whereas BayesA achieved the lowest prediction ability of r = 0.75.

Table 4. 5: Prediction performance of parametric models developed from BGLR models for beta-carotene content prediction before and after adding GWAS significant SNPs as fixed effects, and the percentage improvement of the models after adding significant SNPs. The values are correlation coefficients between predicted and actual values

	Model 1		**Model 2a**		**Model 2b**		**% Improvement**	
	Train.	**Valid.**	**Train.**	**Valid.**	**Train.**	**Valid.**	**Model2a**	**Model 2b**
BRR	0.99	0.14	0.95	0.41	0.93	0.78	193%	457%
BL	0.99	0.14	0.87	0.43	0.88	0.81	207%	479%
BayesA	0.99	0.16	0.98	0.42	0.98	0.75	163%	350%
BayesB	0.99	0.14	0.98	0.41	0.95	0.77	193%	450%
BayesC	0.99	0.15	0.94	0.42	0.94	0.78	167%	420%

Model1 = Initial GP model, Model 2a = GP model after adding 2 significant SNPs above Bonferroni threshold as fixed effect, Model 2b = GP model after adding 5 significant SNP above less stringent threshold of LOD =3 as fixed effect, Train. = Prediction ability of the training set, Valid. = Prediction ability of validation set

4.3.3.2 Non-Parametric GP models

Prediction ability of non-parametric ML models without variable selection is presented in Table 4.6. Based on the prediction ability on the validation set, the models yielded a highest prediction ability (r = 0.44) resulting from the ensemble and the lowest (r= 0.04) resulting from SVMRadial. Similar trend was observed when these models predicted the training set where the ensemble yielded the highest prediction ability (r = 0.37) whereas SVMRadial resulted in the lowest prediction ability (r = 0.09)

Feature selection resulted to 13 SNPs out of 37,419 SNPs as the most informative variables to use for further model development (Figure 4.3). As indicated in Table 4.6, the prediction ability of

non-parametric GP models improved when the 13 important features (SNPs) were used to fit the ML models.

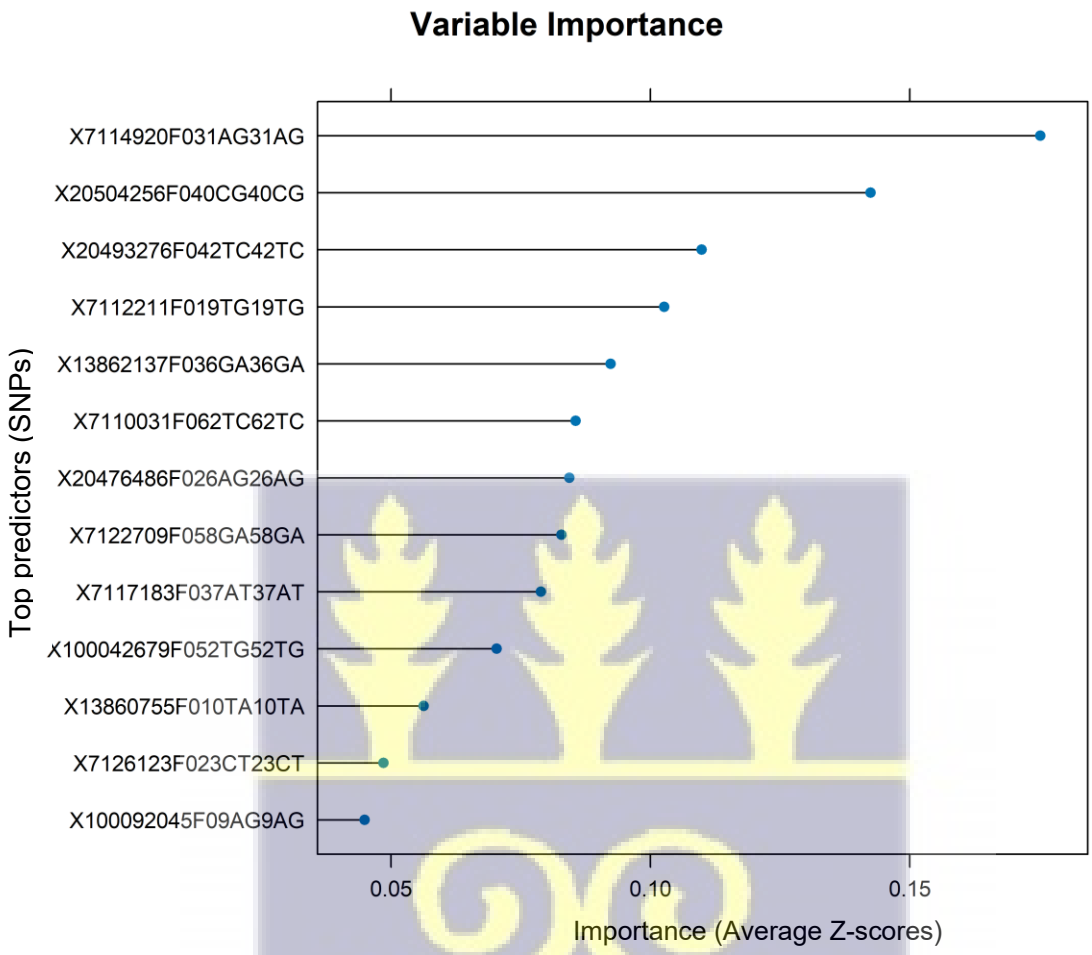


Figure 4. 3: Boruta variable importance plot showing the relative contribution of the top predictors (SNP markers) to the model's performance

From the predictions on the validation set using models developed with feature selection, RF had the highest model prediction ability of r = 0.79 followed by XGBOOST and KNN both having r = 0.73. The lowest prediction ability was presented by SVM and NNET with r = 0.38 and r = 0.34 respectively (Table 4.6). The ensemble of these models resulted in the highest prediction

ability (r = 0.79) equal to that of RF. Similarly, predicting the training population resulted in RF and ensemble having the highest predictive ability (r = 0.72) compared to the rest.

Table 4. 6: Prediction Performance of non-parametric genomic prediction models developed from machine learning algorithms for beta-carotene. The values are correlation coefficients between predicted and actual values

	GP model without variable selection		GP model with variable selection	
	Training	**Validation**	**Training**	**Validation**
RF	0.36	0.21	0.72	0.79
SVMRadial	0.09	0.04	0.40	0.38
NNET	0.25	0.18	0.57	0.34
XGBTree	0.25	0.20	0.54	0.73
KNN	0.16	0.32	0.62	0.73
Ensemble	0.37	0.44	0.72	0.79

4.4. Discussion

The pVAC cassava genetic resource had beta-carotene content in cassava roots that ranged between 0.04 -1.03 Mg/100g (0.40-10.30µg/g). This shows that some genotypes achieve just 1µg retinol activity equivalent (RAE) which is suggested by Institute of Medicine Food and Nutrition (2001) to be equal to 6 µg/g beta-carotene. An adult of 19-50 years old male requires 900 µg RAE while the female requires 700 µg RAE (National Institutes of Health, 2023). Despite the current germplasm not containing significantly high beta-carotene in cassava root to achieve the dietary requirement, it provides a source of pVAC alleles that can be used to enhance the population reach the required level of pVAC.

The 94 pVAC cassava genotypes used in this study had some level of relatedness as shown in the genetic relationship heatmap. This is important to the study because the accuracy of GP model is affected by the extent of genetic relatedness between training population and validation/testing population (Desta & Ortiz, 2014). Since the genotypes in the population had some level of genetic

relatedness, the training and validation sets were related. After quality control, 37,419 SNPs were retained for use in developing GP models. Marker density contributes to the accuracy of GP models especially in the population that has quick LD decay (Ning et al., 2022). The retained SNP markers for modelling were dense enough to achieve reliable parametric GP models. Indeed, high-density markers increase the chance of getting sufficient informative factors (variable importance) to use in non-parametric ML models.

Addition of significant SNPs as fixed effect to parametric BLGR models improved the prediction ability of the final model in predicting cassava root flesh colour. When the single significant SNP on chromosome 01 from multi-locus fixed marker effect GWAS was added to the BGLR models, the percentage improvement ranged from 27% for BL to 62% for BRR models. Similar results were reported by Esuma et al. (2021) who found that GP models improved when significant SNPs associated with total carotenoids content on chromosome 1 were added to the model as fixed effect.

When two significant SNPs on chromosomes 09 and 14 from multi-locus fixed marker effect GWAS were added to basic parametric BGLR models as fixed effects, the percentage improvement of the models ranged between 163% for BayesA and 207% for BL models. This improvement is approximately triple the improvement of the GP model when one significant SNP on chromosome 01 was added. Furthermore, when the five significant SNPs identified by multi-locus random marker effect GWAS on chromosomes 01, 03, 04, 14 and 18 were added to the GP models as fixed effects, the percentage improvement of the models ranged between 350% for BayesA and 479% for BL models. This result showed a trend where GP model improves as the number of significant SNPs from different chromosomes increase. This could be attributed to the fact that more SNPs across the chromosomes could be linked to all causative genes thus capturing all the effects associated with the trait.

The final parametric GP models had high prediction ability that ranged between r = 0.75 to r =0.81. These prediction abilities are reliable enough in predicting GEBV of pVAC cassava. The GEBVs are additive genetic effects that are useful for timely selection of progenitors. In breeding, progenitors are only able to transmit additive genetic effects to the offspring and thus, individuals with high GEBV are always selected. Knowledge of an individual's additive genetic effects at early stages facilitates early selection of parents and this shortens breeding cycle time, by increasing the rate of genetic gains.

Looking at the basic (non-optimized) models' prediction ability, parametric models had higher prediction ability (r = 0.96 - 0.99) on the training sets (genotypes used in model development) and a low prediction ability on the validation set (r = 0.14 – 0.16). On the other side, non-parametric models resulted in low predictions values that ranged from r = 0.09 to r = 0.36 and r = 0.04 to r = 0.44, respectively on training and validation sets. Therefore, parametric models had a huge gap between training set predictions and validation set predictions, and this was not the case with non-parametric models. This can be attributed to the fact that parametric models fit simpler assumptions and memorize training data better, resulting in good predictions in the training set but poor generalization while non-parametric models avoid assumptions and capture complexity, resulting in a small difference in prediction ability between training and testing sets (Imam et al., 2024). This result implies that the basic parametric models had overfitting compared to non-parametric models. Overfitting is observed when the model results to high prediction of training set but extremely low prediction of testing. Nevertheless, both parametric and non-parametric models resulted in good predictions after optimization.

Non-parametric GP models developed without variable selection resulted in a relatively poor model prediction ability. Based on prediction of the validation set, these models had their highest

prediction ability (r = 0.44) resulting from ensemble followed with KNN with r = 0.32 and SVMRadial resulting to the lowest prediction ability (r =0.04). In prediction of the training set, ensemble had the highest prediction ability (r= 0.37) whereas SVMRadial had the lowest performance (r = 0.09).

Variable or feature selection for ML models reduced the number of predictors (SNPs) from 37,419 to just 13. Variable selection is the extraction of most useful model predictors and discarding redundant predictors without affecting model prediction accuracy (El Touati et al., 2024). Variable selection reduces data complexity while improving model focus, which has been reported to improve the accuracy of ML models and reduce the computation time (Aalborg et al., 2024; Touati et al., 2024).

The prediction ability of three ML models; RF, XGBOOST and KNN in predicting beta-carotene content was r = 0.79, r = 0.71 and r =0.71, respectively. The ensemble of these models resulted to the highest prediction ability (r = 0.79) similar to that of the highest preforming single model (RF). In the two cases (one with variable selection and the other without) of GP modeling using machine learning techniques, the ensemble model outperformed all other models in terms of prediction performance. An ensemble model, or enhanced model, is produced by combining the predictions of multiple basic models. While one individual model (RF) yielded a comparable prediction performance of r = 0.79, the ensemble produced a better model overall. This suggests, in general, that although ensembles improve ML models' capacity for prediction, individual models can nevertheless occasionally generate extremely accurate predictions.

4.5. Conclusion

This study compliments the current effort of addressing VAD among cassava consuming communities in Kenya. The developed parametric models were reliable to predict breeding values for pVAC cassava demonstrating their suitability in rapid genetic improvement of pVAC cassava. Adding significant SNPs from GWAS tremendously improved the BGLR genomic prediction models. The results showed that the GP models improved as more significant SNPs across the 18 cassava chromosomes were added as fixed effects in the GP models. Multi-locus random marker effect GWAS has a modified threshold of p-value and high statistical power, making it possible to get more causative SNPs across the cassava genome.

Non-parametric ML models were found to be reliable in predicting genomic value of a genotype by accommodating both additive and non-additive gene effects. Feature selection and ensemble significantly improved model computation time and enhanced prediction power of ML models. RF, XGBOOST, KNN and ensemble of the models were highly accurate and thus, they can be used for final selection of promising cassava genotype.

CHAPTER FIVE

5.0 DETERMINATION OF GENOMIC REGIONS, SUPERIOR ALLELES AND GENES REGULATING BETA-CAROTENE IN CASSAVA ROOTS

5.1 Introduction

Carotenoids are widely recognized for their nutritional benefits, especially in preventing various eye conditions and human cancers (Crupi et al., 2023). With its high vitamin A activity, beta-carotene is the most significant pro-vitamin A carotenoid (Rodriguez-amaya & Kimura, 2004) . Recent years have seen the use of genome-wide association studies (GWAS) to identify putative genetic areas linked to different plant traits. The main purpose of GWAS is to identify quantitative trait nucleotides (QTNs) for use in mining superior or favorable marker alleles, and determining of candidate genes that are around the QTNs. The QTNs identified from GWAS are used to develop KASP markers for use in marker-assisted selection (MAS).

Modern breeding methods including marker-assisted selection (MAS) and genomic selection (GS) can be used to accelerate genetic improvement particularly by reducing generational interval and increasing selection intensity (Ceballos et al., 2015; Ceballos et al., 2020; Mbanjo et al., 2021). However, integration of molecular markers as part of MAS in breeding pipelines requires an initial investment in discovery research to identify major-effect loci that serve as the targets of selection. With the rapid advances in next-generation sequencing (NGS) technologies, it is now feasible to generate genome-wide marker data in large populations. This, coupled with phenotype data makes it possible to identify and map quantitative trait loci (QTL) and genes that are of agricultural importance at the whole genome level (Ikeogu et al., 2019).

Several genome-wide association studies (GWAS) have been conducted to describe the genetic architecture of a number of cassava traits including carotenoids content (Mangal et al., 2024).

These works were carried out on single locus GWAS (SL-GWAS) models such as GLM and MLM. The stringent boniferroni p-value correction in SL-GWAS has been reported to cause exclusion of important loci associated with traits (Zhang et al., 2020). Furthermore, large experiment errors typical in field experiments contribute to SL-GWAS excluding important trait loci (Zhang et al., 2019). This necessitates the need for extensive sample sizes to minimize experimental error, and this is far beyond the reach of resource poor research institutes. Therefore, it is thought that other loci controlling pVAC in cassava genome have not been identified. To address this issue, a multi locus GWAS (ML-GWAS) such as multi-locus random SNP effect mixed linear model (mrMLM) has been suggested (Wang et al., 2016; Zhang et al., 2019; Zhang et al., 2020).

Compared to SL-GWAS, the ML- GWAS has higher statistical power; leading to discovering more QTNs from marker-trait association in small populations (Wang et al., 2016). The mrMLM has two stages of analysis: first, all SNPs are treated as random effect and then algorithms are used to select potential significant SNPs (Wang et al., 2016; Zhang et al., 2020). Under this random model, the estimated marker effects are shrunk towards zero. This shrinkage makes the majority of the markers (having zero effects) to be eliminated. In the second stage, all selected SNPs are fitted into one model and those that pass through the modified Bonferroni correction (p-value = $0.05/M_e$, where M_e is the effective number of markers) are further screened through likelihood ratio test (LRT) for true QTN (Zhang et al., 2020). Markers in the LRT that surpasses the LOD score threshold are regarded to be significantly associated with the trait under study (Zhao et al., 2022).

According to recent research, GWAS analysis incorporating haplotypes can outperform traditional GWAS analysis in terms of allelic effect estimation and statistical significance (better p-values) (Sehgal et al., 2020). One major drawback of SNPs is that they only offer bi-allelic information at

each individual locus; they do not contain the polymorphic information content (PIC) needed to identify genetic diversity for a particular trait (Lu et al., 2011). Variants with a minor allele frequency of less than 0.05 and markers with a missing rate of more than 10% are generally not taken into account by GWAS (Tabangin et al., 2009). Because haplotype-based GWAS are typically multi-allelic and have better LD than individual SNPs, they are able to circumvent this constraint by increasing the allelic resolution of variations (Bajgain & Anderson, 2021; Lisker et al., 2022) .

Recent years have seen the development of two haplotype-based GWAS approaches that have demonstrated increases in mapping power over SNP-based methods in plant datasets: Reliable Association INference By Optimizing Weights (RAINBOW) (Hamazaki & Iwata, 2020) and Functional Haplotype GWAS (FH-GWAS) (Liu et al., 2019). These investigations have shown that employing haplotypes to get around problems in plant GWAS is feasible. Because these haplotype-based methods concentrate on the complete haplotype block rather than on each SNP inside the haplotype block, they are able to control false positives more effectively than the single-SNP method (Hamazaki & Iwata, 2020).

For effective deployment of MAS in cassava biofortification with pVAC, there is need to discover all genomic regions associated with pVAC on the cassava genome. Despite the importance of multi-locus fixed marker effect, multi-locus random marker effect and haplotype-based GWAS models, no research has been reported on their use in determining the genetic architecture of pVAC in cassava roots. With its high statistical power and modified p-value threshold, ML-GWAS could discover more loci associated with the trait. Similarly, deploying haplotype-based GWAS may lead to detection of additional loci for pVAC in cassava roots.

This study employed 94 pVAC cassava genotypes that were genotyped using DArT markers and phenotyped for beta-carotene content to:

i) identify genomic regions controlling pVAC in cassava using multi-locus fixed marker effects, multi-locus-random marker effect and haplotype-based GWAS models;

ii) determine SNP marker superior alleles for beta-carotene content; and

iii) determine the candidate genes associated with pVAC in cassava.

5.2 Materials and Methods

5.2.1 Genetic Resources, Phenotyping, Genotyping and SNP quality control

The genetic resource used in this study has been described in section 3.2.1 and genotyped as described in section 3.2.2. Phenotyping of beta-carotene content of the genetic resources has been described in section 4.2.2.

5.2.2 Identification of Genomic Regions and favorable alleles for beta-carotene in cassava

Genomic regions associated with expression of pVAC in cassava root was determined using GWAS. A multi-locus fixed marker effect GWAS was carried out using BLINK model in GAPIT v3.1.0 package in R statistical software. Based on the population structure results (section 3.3.3), three principal components were used in the model to correct against false positives that could arise due to population structure. This analysis was carried out using the following model:

$$\boldsymbol{y} = \boldsymbol{M\mu} + \boldsymbol{\chi_1\beta_1} + \boldsymbol{\varepsilon},$$

Where; **y** represents phenotype vector of beta-carotene content, **M** represents incidence matrix for fixed effects, **μ** represents a vector of fixed effects coefficients (marker effects), χ_1 represents

incidence matrix of random effect, $\boldsymbol{\beta_1}$ represents random vector of population structure: $\boldsymbol{\beta_1 \sim N(0, I\sigma^2_{\beta 1})}$, $\boldsymbol{\varepsilon}$ represents vector of residual terms: $\boldsymbol{\varepsilon \sim N(0, I\sigma^2_{\varepsilon})}$.

Subsequently, a multi-locus random marker effect GWAS was carried out using four models (mrMLM, FASTmrMLM, pLARmEB and ISIS EM-BLASSO) in mrMLM v4.0.2 package in R statistical software. Population structure was accounted for by use of three principal component. The critical threshold of the LOD score for SNP marker effects was set at 3.

Haplotype-based GWAS was carried out using haplotype blocks identified on the cassava genome. Solid Spine of LD method was used to identify the haplotype blocks. In this analysis, LD matrix was constructed from SNP markers using *snpStats* package in R. Solid Spine of LD method requires that the first, last and intermediate SNPs of a block be at a strong LD (Kim et al., 2018). Therefore, haplotype blocks were identified at an LD threshold of R^2=0.80 using a function developed using basic R. The identified haplotype block were visualized through plots developed using *patchwork* and *ggplot2* packages in R. The haplotype-based GWAS was carried out using Reliable Association INference By Optimizing Weights with R (RAINBOWR) package in R where haplotype effects were tested using the score test method. RAINBOWR provides likelihood-ratio test and score test options for testing haplotypes effects. In this study, score test method was adopted because it is more faster than likelihood-ratio test. Linear kernel method, additive test effect and three principal components (fixed effect to account for population structure) were used. Haplotype-based GWAS model was fitted using the following model:

$$\boldsymbol{y = X\beta + Z_h h + Z_g g + \varepsilon}$$

Where **y** is a vector of phenotype (beta-carotene), **X** is incidence matrix of fixed effects, **β** is vector of fixed effects (population structure, covariates), $\mathbf{Z_h}$ is incidence matrix for haplotype effects, $\boldsymbol{h}$

is a vector of haplotype effects, $\mathbf{Z_g}$ is an incidence matrix for genetic background effects (polygenic effects), **g** is a vector of polygenic effects: $\mathbf{g \sim N(0, G\sigma_g^2)}$, **G**=genomic relationship matrix and **ε** is the vector of residual terms: $\boldsymbol{\varepsilon \sim N(0, I\sigma_\varepsilon^2)}$.

CMplot package in R was used to visualize the GWAS result using manhattan plot and QQ-plot. Favorable SNP alleles for beta-carotene were identified by plotting allele genotypes of significant SNP loci against beta-carotene values. Alleles associated with high beta-carotene values were selected from the genotypes in the plot reflecting a higher median value and termed as favorable alleles

5.2.3 Candidate gene Identification and Functional Analysis

Candidate genes were screened using the *Manihot esculenta* v8.1 reference genome, which was retrieved from the *Phytozome* version 13 database using the *biomaRt* package in R. Based on the extent of LD decay identified in the population at R^2=0.20, candidate genes were screened on this reference genome within a window of 613.072 kb upstream and downstream of significant SNPs. Generally, an LD threshold of R^2=0.10 and R^2=0.20 are used in LD decay analyses. In this study, an LD threshold of R^2 = 0.20 was used to set the LD decay distance. This LD threshold provides a short window (LD decay distance) for screening candidate genes to ensure that they are strongly linked to the significant SNPs. The *ClusterProfiler* package in R was used to conduct gene ontology analysis based on biological processes (BP) involved by the candidate genes. Background genes for *Manihot esculenta* were obtained from the *OrgDb* annotation package using the *AnnotationHub* package in R.

5.3 Results

5.3.1 Marker-trait association

GWAS results from the BLINK model identified two significant SNPs associated with beta-carotene on chromosomes 09 and 14 respectively at position 34394428 and 6077652. Five significant SNPs located on chromosomes 01, 03, 04, 14, and 18 significantly linked to beta-carotene in cassava roots were detected using multi-locus random marker effect GWAS models implemented in mrMLM. The QQ-plots for all GWAS models used, exhibited reliable marker trait association (MTA) resulting from the analyses (Figure 5.1).

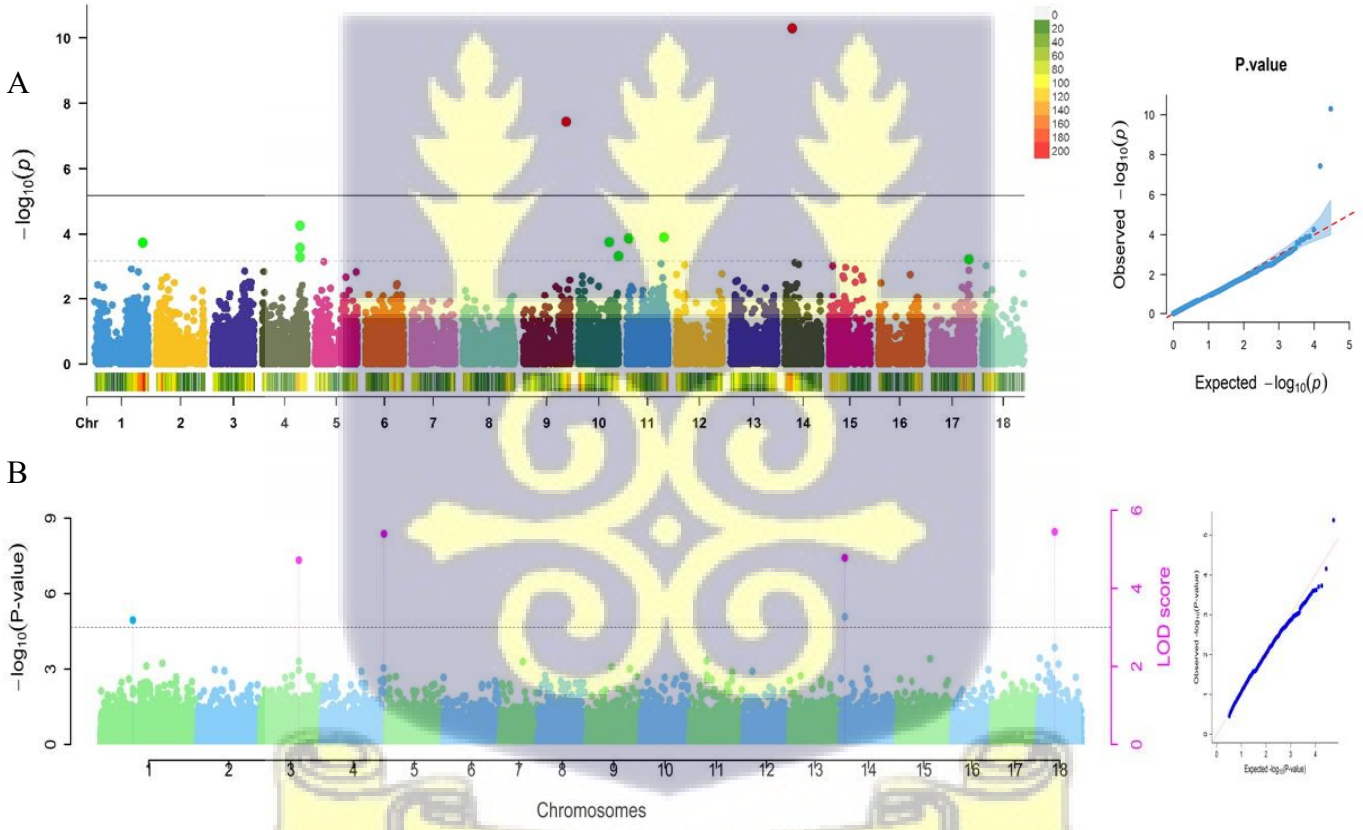


Figure 5. 1: Manhattan and QQ-plots for GWAS models: A= Manhattan plot from Blink GWAS model and B= Manhattan plot from multi locus random marker effect GWAS models (mrMLM). In the mrMLM models' Manhattan plot, the purple hits are significant SNPs identified by more than one model of multi locus random marker effect GWAS, while the blue hits are significant SNPs detected by one model

One notable SNP is that on chromosome 14, position 6077652, which was identified by all models to be significantly associated with beta-carotene (Table 5.1). Other SNPs that were identified by two GWAS models were those on chromosomes 03, 04 and 18 respectively at position 5342366, 7941731 and 24415388. The significant SNPs on chromosomes 01, 03 and 14 were as a result of transition mutations. On the other hand, SNPs on chromosomes 04, 09 and 18 resulted from transversion mutations.

Table 5. 1: Significant SNPs from GWAS models for beta-carotene content in cassava roots

GWAS model	Chromosome	Position (bp)	Ref/Alt
1	01	5355296	T/C
2 &3	03	5342366	A/G
4 & 1	04	7941731	C/A
5	09	34394428	T/A
1,2,3,4 & 5	14	6077652	A/G
1 & 4	18	24415388	C/G

bp = Base pairs, 1= mrMLM, 2= pLARmEB, 3 = FASTmrMLM, 4 = ISIS EM-BLASSO and 5 = BLINK. RS#= Reference SNP number, QTN = Quantitative trait nucleotide

Figure 5.2 shows the distribution of SNP markers that were clustered into haplotype blocks on each chromosome of the cassava genome. The result showed that from the large dataset of SNPs (as shown on Table 3.3), the number of SNP markers clustering into haplotype blocks reduced to few hundred across all 18 chromosomes of cassava. Chromosome 04 had the highest number of SNPs (333) forming haplotype blocks while chromosome 07 had the least number of SNPs (98).

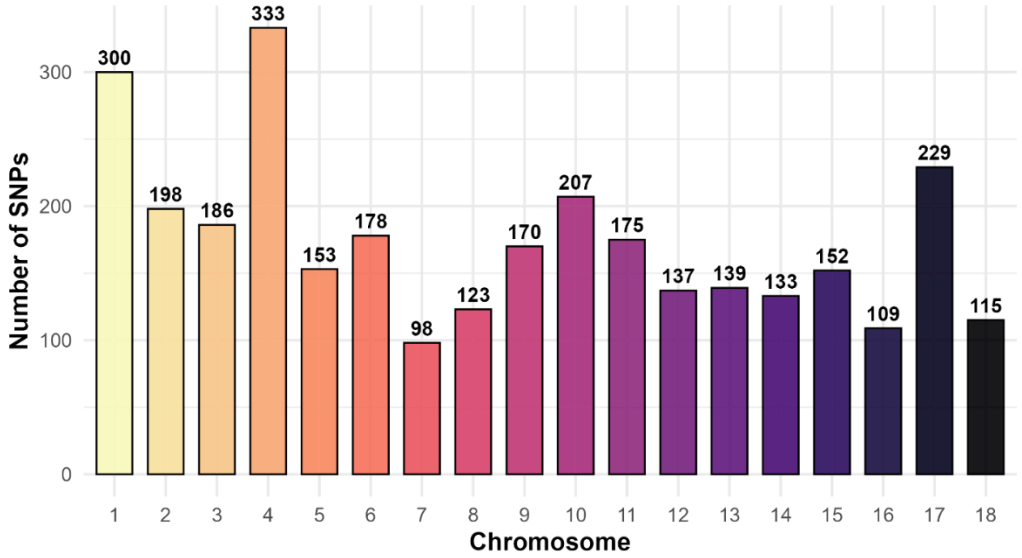


Figure 5. 3: Distribution of SNPs forming haplotype blocks in the cassava genome

Distribution of SNP markers in each haplotype block is presented on Figure 5.3. In this result, the highest number of SNPs per haplotype block was 6. About 21 haplotype blocks each having 4 - 6 SNP markers were observed. Majority of the region on the cassava genome had short haplotype blocks each with 2-3 SNP markers and this appeared to be the lowest markers distribution.

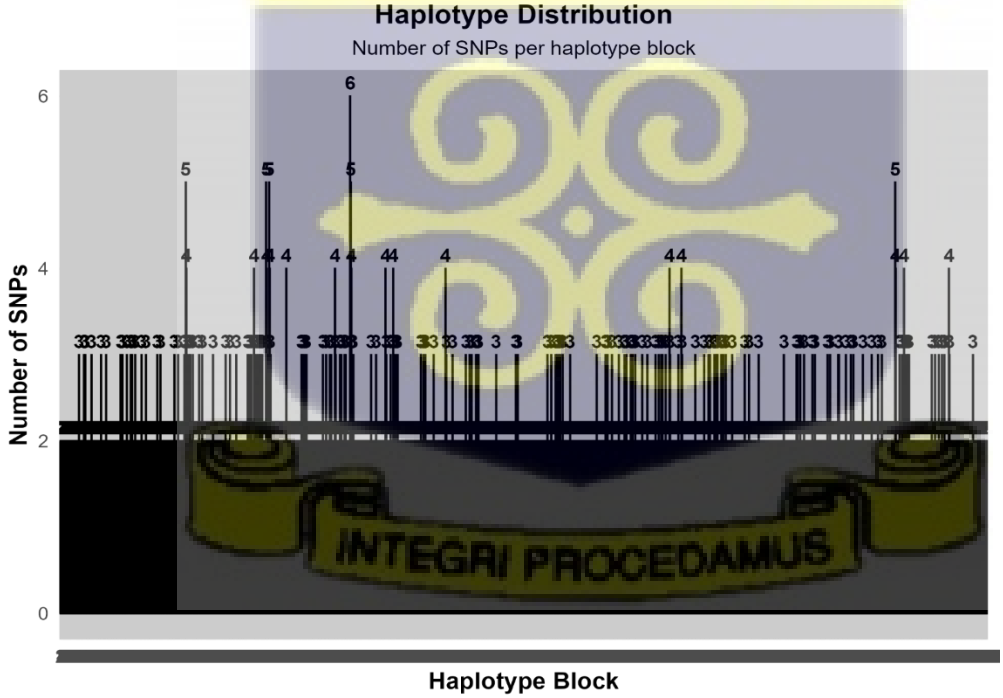


Figure 5. 4: Marker distribution on each haplotype block

Figure 5.4 shows the distribution of haplotype blocks on each chromosome of the cassava genome. The result showed that there was a total of 1,468 haplotype blocks on the cassava genome distributed on all 18 chromosomes of cassava. Chromosome 04 had the highest number of haplotype blocks (146) while chromosome 07 had the lowest number of haplotype blocks (46). All other chromosomes apart from chromosomes 07, 16 and 18 had more than 60 haplotype blocks.

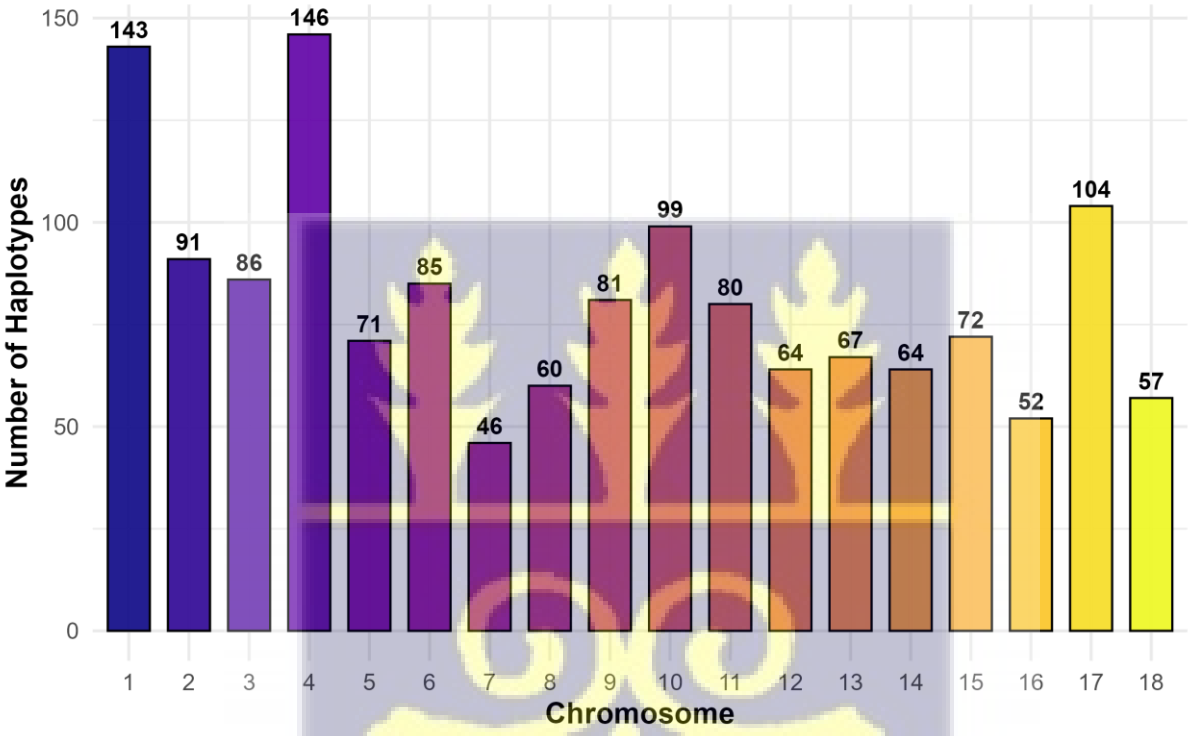


Figure 5. 5: Distribution of haplotype blocks in each chromosome of cassava genome

Fifteen haplotype blocks on 10 chromosomes were observed to have a significant association with beta-carotene (Figure 5.5). Chromosomes 01, 03 and 04 had haplotype blocks exhibiting very high and significant effects (-log10 (p) > 6.0) for beta carotene among other significant haplotype blocks.

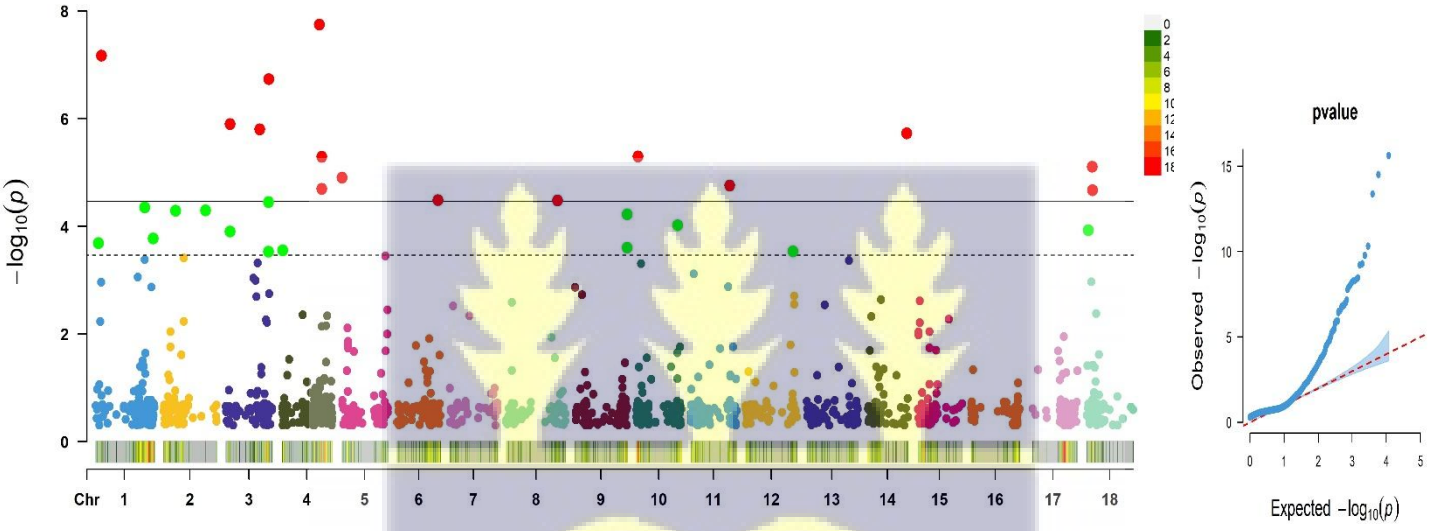


Figure 5. 8: Manhattan and QQ-plot for haplotype-based GWAS

Results indicate that ten chromosomes: 01, 03, 04, 05, 06, 08, 10 11, 14 and 18 had significant haplotype blocks for beta-carotene (Table 5.2). Out of 1,468 haplotype blocks, 15 blocks were located on regions that are significantly associated with beta-carotene. Chromosomes 03 and 04 each had three haplotype blocks located on regions that are significantly associated with the trait. Similarly, chromosome 18 had two haplotype blocks located on regions associated with the traits.

Table 5. 2: Genomic coordinates and significance values of haplotype blocks that were in significant association with beta-carotene in cassava roots

	Haplotype Block	Chromosome	Position	p-Value
1	57	01	5443919	6.73E-08
2	256	03	30374336	1.86E-07
3	244	03	3220936	1.27E-06
4	306	03	24132836	1.58E-06
5	398	04	26902638	1.8E-08
6	444	04	28509293	5.1E-06
7	445	04	28509293	2.01E-05
8	523	05	790785	1.24E-05
9	586	06	28859091	3.29E-05
10	718	08	35791856	3.31E-05
11	887	10	1538515	5.06E-06
12	916	11	27900387	1.75E-05
13	1159	14	28153367	1.88E-06
14	1412	18	4175031	7.78E-06
15	1425	18	4304262	2.13E-05

Table 5.3 shows results on the ability of different GWAS models in detection of SNPs that are significantly associated with beta-carotene in cassava roots. Among the studied GWAS methods a haplotype based GWAS method implemented using RAINBOWR model detected the highest number (15) of significant marker-trait associations. Multi-locus random marker effect GWAS models had disparity in detection of significant SNPs. This follows the results showing that 4, 3, 3 and 2 SNPs, respectively, for mrMLM, ISIS EM-BLASSO, pLARmEB and FASTmrMLM were in significant association with beta-carotene content. Multi-locus fixed marker effect GWAS implemented in BLINK model detected 2 significant SNPs.

Table 5. 3: Comparison of different GWAS models in detection of causative SNPs

GWAS Method	GWAS Model	Significant SNPS
Multi-locus fixed marker effect	BLINK	2
Multi-locus random marker effect	mrMLM	4
	pLARmEB	2
	FASTmrMLM	2
	ISIS EM-BLASSO	3
Haplotype based GWAS	RAINBOWR	15

5.3.2. Identification of favorable alleles

Table 5.4 show the effect of alternative alleles in relation to reference allele resulting from GWAS analysis. This result showed that the alternative alleles C and A respectively for SNPs in chromosomes 01 and 09 had a positive beta carotene effect in relation to reference alleles. The remaining chromosomes had alternative alleles G, A, G and G respectively on chromosomes 03,04, 14 and 18 having negative beta carotene effects.

Table 5. 4: Allele effects of SNPs that were significantly associated with beta carotene

Chromosome	Position (bp)	Ref/Alt	Allele effect
01	5355296	T/C	0.116
03	5342366	A/G	-0.144
04	7941731	C/A	-0.168
09	34394428	T/A	0.214
14	6077652	A/G	-0.285
18	24415388	C/G	-0.281

Phenotypic contribution of genotypes from the six significant SNPs for beta-carotene content in cassava root is shown in Figure 5.6. The number of genotypes for each SNP marker varied from three to one. For instance, SNPs on chromosome 01, 03, 09 and 14 had two genotypes; while those on chromosomes 04 and 18, respectively, had three genotypes and one genotype. In this study,

SNP alleles that were associated with increased beta-carotene content were defined as favorable alleles. These alleles were identified by comparing mean phenotype values across genotypes, where alleles in genotypes with the highest means were considered favorable. Alleles C, G, C, T and G, respectively, for SNP markers on chromosomes 01, 03, 04, 09 and 14 exhibited a higher beta-carotene level in cassava roots and thus, selected as favorable or superior alleles. The SNP on chromosome 18 has one genotype with homozygous allele and this indicated that the loci is fixed.

.

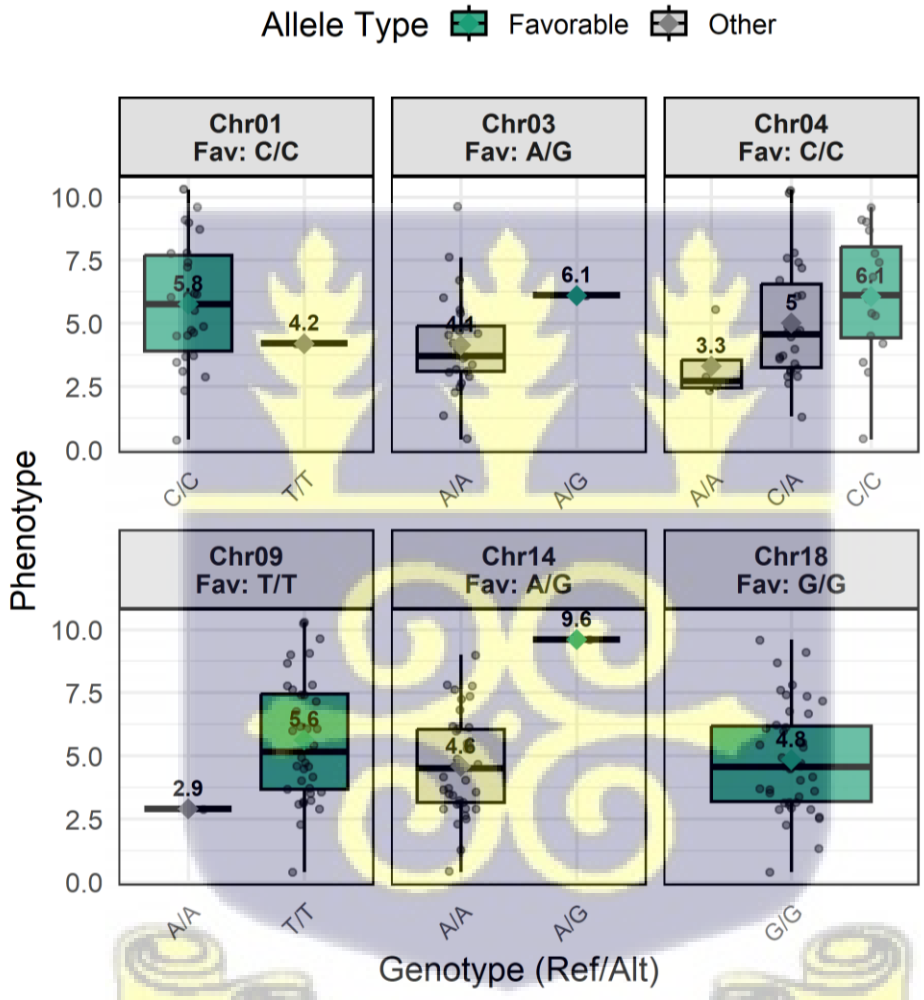


Figure 5. 9: Phenotypic contribution of significant SNPs on chromosomes01-position 5355296, chromosome03- position 5342366, chromosome04-position 7941731, chromosome09-position 34394428, chromosome14-position 6077652 and chromosome18-position 24415388, Fav = Favorable alleles.

5.3.3. Candidate gene identification

Three candidate genes; *Manes.01G124200, Manes.09G079657* and *Manes.03G084700* were found to be involved in carotenoids biosynthesis in cassava (Table 5.5). Functional analysis of these three genes identified *Manes.01G124200* gene (phytoene synthase) which is involved in carotenoid biosynthesis in cassava at the vicinity of chromosome 01 SNP. Similarly, *Manes.03G084700* gene, which is also a Phytoene synthase gene, was identified around the significant SNP on chromosome 03. Candidate gene, *Manes.09G079657* (CYTOCHROME P450 97B3, CHLOROPLASTIC) which is involved in carotenoids biosynthesis in cassava was identified to be located at the vicinity of the SNP on chromosome 09. Functional analysis of candidate genes associated with the remaining SNPs revealed that they are involved in carotenoids catabolism activities such as apocarotenoids biosynthesis leading to production of abscisic acid. According to gene ontology (GO) analysis of biological process (BP), the candidate genes near SNPs on chromosomes 14 and 18 were involved in apocarotenoid biosynthesis. However, the GO IDs were not matching to any known gene in cassava.

Table 5. 5: Details of the candidate genes involved in beta-carotene biosynthesis pathway

Chr	SNP Pos. (bp)	Gene ID	Gene name	Function	GO ID (Biological Process)
1	5355523	Manes.01G124200	phytoene synthase (crtB)	Carotenoid biosynthesis	GO:0009058, GO:0016740
3	5342366	Manes.03G058000 Manes.03G057900	BETA-CAROTENE DIOXYGENASE	Abscisic acid biosynthesis	Not available
		Manes.03G057750	carotenoid cleavage dioxygenase (K11159)	Abscisic acid biosynthesis	Not available
		Manes.03G084700	phytoene synthase (crtB)	Carotenoid biosynthesis	GO:0009058, GO:0016740
4	7941731	Manes.04G055244	Aldehyde oxidase / Retinal oxidase	Carotenoid catabolism (Abscisic acid biosynthesis)	

9	34394428	Manes.09G079657	CYTOCHROME P450 97B3, CHLOROPLASTIC	Carotenoid biosynthesis	GO:0004497, GO:0005506, GO:0016705, GO:0020037, GO:0055114
		Manes.09G144600	Violaxanthin de-epoxidase (VDE, NPQ1)	Carotenoid biosynthesis	GO:0009507, GO:0046422, GO:0055114
14	6077652	Not available	Not available	Apocarotenoid biosynthesis (Abscisic acid biosynthesis)	GO:0043288, GO:0043289
18	24415388	Not available	Not available	Apocarotenoid biosynthetic	GO:0043289, GO:0009688

Chr = Chromosome, GO =Gene Ontology

Table 5.6 presents functional analysis results of the new candidate genes identified by haplotype-based GWAS. On top of these new candidate genes, all other genes identified using the results from multi-locus fixed marker effect GWAS and multi-locus random marker effect GWAS were as well identified by haplotype-based GWAS. Detailed results from the analysis of candidate genes that were in close proximity with genomic regions detected by haplotype-based GWAS model for pVAC in cassava is presented in Appendix 1.

Among the candidate genes identified through haplotype-based GWAS results, new genes included *Manes.01G001200* (Zeta-carotene isomerase / 15-cis-zeta-carotene isomerase) on chromosome 01 which is responsible for carotenoid biosynthesis (Table 5.5). On chromosome 03, *Manes.03G083500* that is 9-cis-epoxycarotenoid dioxygenase (NCED) and *Manes.03G150400* (BETA-CAROTENE DIOXYGENASE) both involved in abscisic acid biosynthesis were identified. On chromosome 05, *Manes.05G193700* which is a 15-cis-phytoene desaturase (PDS, crtP) that controls carotenoid biosynthesis was identified. Other candidate genes identified on this chromosome were *Manes.05G082900* (beta-carotene isomerase (DWARF27)),

Manes.05G051700 (9-cis-beta-carotene 9',10'-cleaving dioxygenase), and *Manes.05G005000* (Beta-carotene isomerase) all involved in 5-deoxystrigol biosynthesis. On chromosome 08, *Manes.08G037100* which is prolycopene isomerase (crtISO, crtH) and *Manes.08G016300* which is carotene epsilon-monooxygenase (LUT1, CYP97C1) both involved in carotenoid biosynthesis were identified.

Table 5. 6: A portion of unique candidate genes that were identified only by haplotype-based GWAS.

HB	Chr.	Pos.	Gene ID	Gene Name	Function
57	1	5443919	Manes.01G001200	Zeta-carotene isomerase / 15-cis-zeta-carotene isomerase	carotenoid biosynthesis
256	3	30374336	Manes.03G083500	9-cis-epoxycarotenoid dioxygenase (NCED)	abscisic acid biosynthesis
244	3	3220936	Manes.03G150400	BETA-CAROTENE DIOXYGENASE	abscisic acid biosynthesis
523	5	790785	Manes.05G193700	15-cis-phytoene desaturase (PDS, crtP)	carotenoid biosynthesis
			Manes.05G082900	beta-carotene isomerase (DWARF27)	5-deoxystrigol biosynthesis
			Manes.05G051700	9-cis-beta-carotene 9',10'-cleaving dioxygenase	5-deoxystrigol biosynthesis
			Manes.05G005000	Beta-carotene isomerase	biosynthesis of strigolactones
718	8	35791856	Manes.08G037100	prolycopene isomerase (crtISO, crtH)	carotenoid biosynthesis
887	10	1538515	Manes.10G141300	CAROTENOID 9,10(9',10')-CLEAVAGE DIOXYGENASE 1	apocarotenoid biosynthesis
916	11	27900387	Manes.11G097600	Phytoene desaturase (zeta-carotene-forming) / 2-step phytoene desaturase	zeaxanthin biosynthesis
1159	14	28153367	Manes.14G022700	Beta-carotene 3-hydroxylase / Beta-carotene 3,3'-monooxygenase	apocarotenoid biosynthesis

HB= Haplotype block number, Chr.= Chromosome

On chromosome 10, *Manes.10G141300* (CAROTENOID 9, 10 (9',10')-CLEAVAGE DIOXYGENASE 1) that is involved in biosynthesis of apocarotenoids was identified. On

chromosome 11, *Manes.11G097600* which is a Phytoene desaturase (zeta-carotene-forming) / 2-step phytoene desaturase involved in zeaxanthin biosynthesis was identified. On chromosome 14, Manes.*14G022700* which is Beta-carotene 3-hydroxylase / Beta-carotene 3, 3'-monooxygenase involved in carotenoid biosynthesis was identified.

Figure 5.7 shows a physical genome map of 20 candidate genes identified in this study, controlling the biosynthesis of beta carotene in cassava roots. Genes on red strips are involved in anabolism of the trait leading to accumulation of beta-carotene while those on blue strips are involved in catabolism leading to reduced amount of the trait.

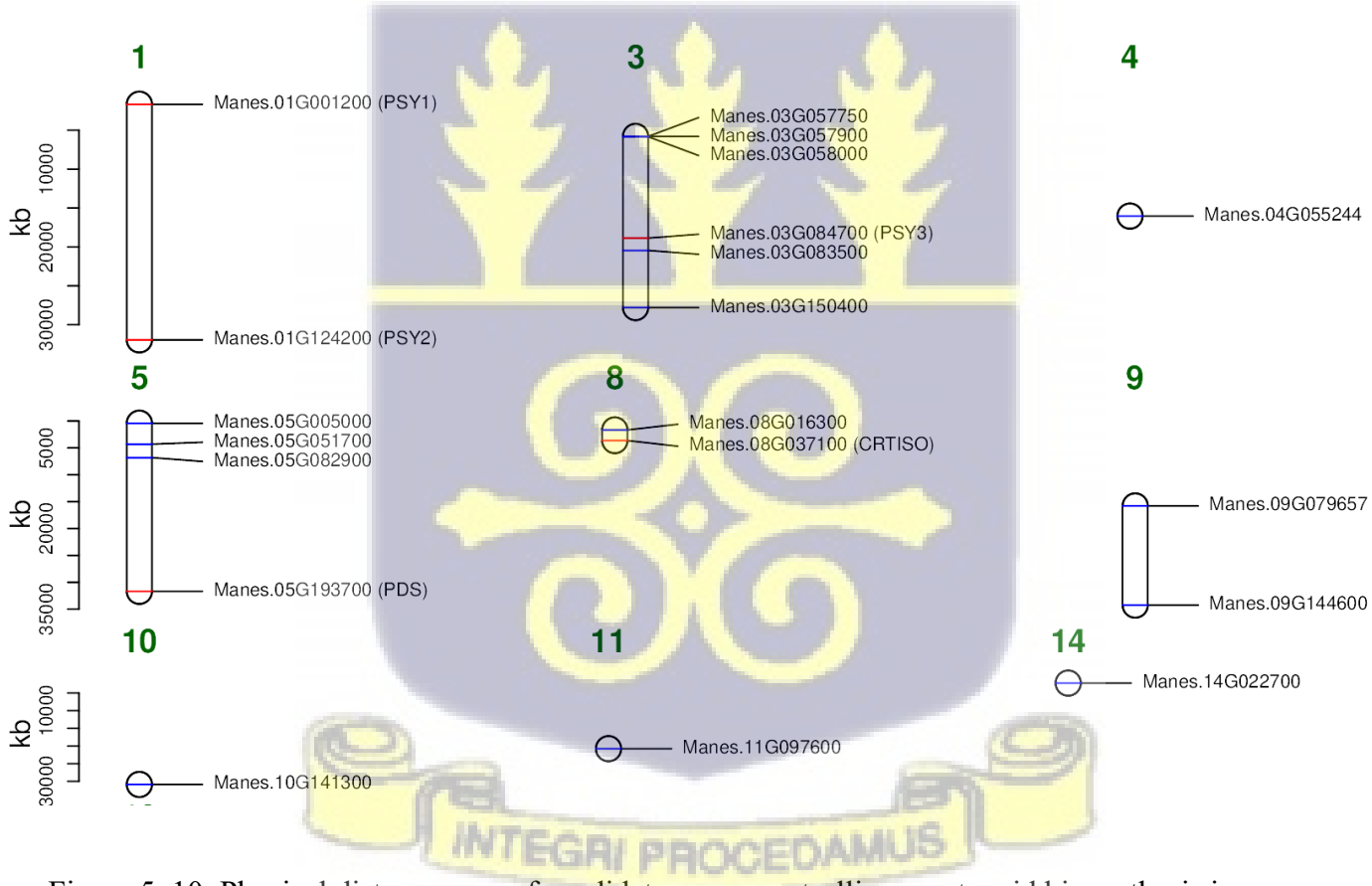


Figure 5. 10: Physical distance map of candidate genes controlling carotenoid biosynthesis in cassava roots

5.4. Discussion

The aim of the present study was to identify genetic loci associated with pVAC in cassava roots using multi-locus random marker effects, and haplotype-based GWAS models. By identifying the loci causing pVAC accumulation in cassava roots, a better understanding on the molecular mechanisms underlying the pVAC accumulation process is achieved. This is vital on the development of effective strategies for genetic improvement of cultivars and makes it easier to identify the genes governing this trait.

Two SNPs that are strongly associated with beta-carotene content of cassava roots on chromosomes 09 and 14 were detected by multi-locus fixed marker effect GWAS using the BLINK model. On the other hand, the multi-locus random marker effect GWAS detected five distinct SNPs linked to beta-carotene using four models on the mr-MLM package. Results showed that out of the five SNPs, four SNPs on chromosomes 03, 04, 14, and 18 were each detected by at least two distinct GWAS models. The SNP found on chromosome 14 may be quite interesting as, in contrast to other SNPs, all four mrMLM GWAS models detected it. This result is in line with earlier reports where significant SNPs on chromosomes 01, 02, 04,13,14, and 15 have been linked to several pVAC (trans beta-carotene, total carotenoid content, cis beta-carotene, violaxanthin, lutein, phytoene, and alpha-carotene) in cassava roots (Ikeogu et al., 2019; Luo et al., 2018).

Compared to traditional fixed marker effects GWAS, the adoption of multi-locus random marker effect GWAS models, resulted in the identification of the majority of causative SNPs due to its strong statistical power with a modified threshold p-value, (Zhou et al., 2021). Small-effect causal SNPs can be found using this random marker effect GWAS methodology (Wang et al., 2016; Zhang et al., 2020). Five chromosomes with significant SNPs linked to beta-carotene may indicate that multiple genes may be involved in the trait's regulation. These findings align with publications

that indicate, even though main genes affect pVAC, a number of genes with minor effects may also be involved in controlling the trait, according to the quantitative root color variability that is observed (Morillo et al., 2012).

Haplotype-based GWAS identified 15 regions (haplotype blocks) on chromosome 01, 03, 04, 05, 06, 08, 10, 11, 14 and 18 significantly associated with the traits. This is the first study to carry out haplotype-based GWAS for pVAC in cassava roots. Nevertheless, some haplotype blocks controlling the trait were located on common chromosomal regions that have been previously reported to control the trait (Ikeogu et al., 2019; Rabbi et al., 2017b). Therefore, new causative variants on chromosomes 03, 05, 06, 08, 09, 10, 11 and 18 that are strongly associated with pVAC were discovered through this study.

Results revealed that a small number of haplotype blocks (about 21) were each consisting of 4-6 SNP markers. However, majority of haplotype blocks were each consisting of 2 - 3 SNP markers and thus, considered to be short. Long haplotype blocks suggest the presence of minimal recombination and high LD in the population. This is common in populations that have experienced a recent strong selection, or have a significant degree of inbreeding (Lamichhaney, 2019). Presence of small number of long haplotype blocks could be attributed to selection and recent inbreeding of the study population. Indeed, the study population consisted of half-sib families whose parents were partial inbred lines subjected to a second cycle of selection (C_2) for pVAC. Majority of the haplotype blocks were short and this indicated a high extent of recombination events that disrupt LD, which is common in diverse and outbred populations (Lamichhaney, 2019). Short haplotype blocks enhance GWAS resolution by making it easier to pinpoint particular genes with the association signal (Wu et al., 2022). Therefore, the resolution of the present marker-trait association analysis was reliable.

A total of 1,468 haplotype blocks were identified on a whole cassava genome composed of 37,419 SNP markers. Loci along the genome are usually inherited in a block-like structure, with only few recombination hotspots and this defines haplotype blocks (Weber et al., 2023). Therefore, the complexity of genomic data (marker data) can be decreased by combining thousands of SNPs into a few hundred haplotype blocks through powerful statistical tests (Sivabharathi et al., 2024). The presence of haplotype-blocks provides cassava breeders with an opportunity to carryout genomic analyses with high statistical power using the study population.

The haplotype-based GWAS implemented in RAINBOWR model was able to detect more significant SNPs (15) compared to multi-locus random marker effect GWAS and multi-locus fixed marker effect GWAS that detected less than 5 significant SNPs associated with beta carotene. This could be attributed to the fact that SNPs do not cover the entire allelic diversity of a gene whereas haplotype-based GWAS covers a wide allelic diversity and gives more accurate results (Abed & Belzile, 2019; Zhou et al., 2021). Therefore, high detection power of haplotype-based GWAS due to its superior LD and multi allelic resolution compared to conventional GWAS models led to discovery of high number of regions associated with the trait.

The phenotypic contribution of genotypes across six significant SNPs revealed clear allelic effects on beta carotene accumulation in cassava roots. Variation in genotype number per SNP indicated differing levels of polymorphism among loci, with chromosomes 01, 03, 09, and 14 showing two genotypes, chromosome 04 exhibiting three, and chromosome 18 being monomorphic. The monomorphic SNP on chromosome 18, represented by a single homozygous genotype, suggests that the locus is fixed, likely due to past selection or genetic drift.

The discrepancy between negative GWAS allele effects (Table 5.4) and positive genotypic contributions (Figure 5.6) reflects the gap between statistical averages and real biological

interactions. GWAS allele effects estimate marginal averages, while genotype plots show actual phenotypic outcomes. Results revealed that on chromosome 01, a positive allele effect (0.116 for *C*) aligns with the higher median beta-carotene in genotype C/C (5.8) compared with T/T (4.2), indicating an additive or dominant positive effect; hence, C is the favorable allele**.** On chromosome 03, despite a negative allele effect (-0.144 for *G*), the A/G genotype shows a higher median (6.1 against 4.4 in A/A), suggesting dominance or epistasis; therefore, G is favorable in the heterozygous state**.** On chromosome 04, a negative effect (-0.168 for *A*) matches the trend C/C (6.1) > C/A (5.0) > A/A (3.3), confirming C as the favorable allele through additive action. On chromosome 09, although the *A* allele shows a positive effect (0.214), T/T (5.6) outperforms A/A (2.9), implying non-additive interactions or allele mislabeling; hence, T is the favorable allele**.** On chromosome 14**,** a strong negative effect (-0.285 for *G*) contrasts with A/G's highest median (9.6 against 4.6 in A/A), indicating dominance or epistasis where G is favorable in the heterozygous combination**.** On chromosome 18, only G/G (median 4.8) is observed, limiting interpretation; however, the negative effect (-0.281 for *G*) suggests C may be the favorable allele**,** though data are insufficient to confirm.

Overall, these results show that favorable alleles for beta-carotene accumulation are C (Chr01, Chr04), G (Chr03, Chr14), T (Chr09), and possibly C (Chr18). The findings highlight that allele effects from standard GWAS can overlook non-additive interactions such as dominance and epistasis that govern the real genetic contribution to beta-carotene content. The identified favorable alleles can serve as molecular markers for selecting clones with superior beta carotene content. The fixed locus on chromosome 18 may indicate a stable, beneficial allele already established in the population.

Cassava genome (*Manihot esculenta v8.1*) on phytozome was queried to identify annotated genes with 613.072 kb upstream and downstream of significant SNPs. Four candidate genes namely, *Manes.01G124200, Manes.03G084700, Manes.09G079657* and *Manes.09G144600* were identified to be involved in carotenoids biosynthesis in cassava. The candidate gene *Manes.01G124200* was identified at a close proximity with the significant SNP on chromosome 01. Results indicate that the *Manes.01G124200* is a *Phytoene synthase* gene, which is involved in carotenoid biosynthesis. There has been similar reports indicating that this gene located in chromosome 01 of cassava genome control accumulation of pVAC in cassava (Ikeogu et al., 2019; Rabbi et al., 2017a). Our study identified *Manes.03G084700* gene, which is also a *Phytoene synthase* gene located in close proximity with significant SNP on chromosome 03. This is the first study to report a genomic variant in chromosome 03 of cassava genome linked to a *Phytoene synthase* gene. Candidate gene *Manes.09G079657* which is a *CYTOCHROME P450 97B3, CHLOROPLASTIC* gene was identified in chromosome 09. Gene ontology analysis indicate that this gene is involved in carotenoids biosynthesis and thus, associated with accumulation of carotenoids levels. Literature survey revealed that *CYTOCHROME P450 97B3* modulates carotenoids accumulation in plants (Zhang et al., 2024). Similarly, *Manes.09G144600* gene which is a *Violaxanthin de-epoxidase* (VDE, NPQ1) was identified at close proximity with significant SNP on chromosome 09. Results from functional analysis revealed that this gene is involved in carotenoids biosynthesis in cassava. However, literature search further revealed that *Violaxanthin de-epoxidase* is involved in conversion of Violaxanthin (apocarotenoid) back to Zeaxanthin (apocarotenoid) which can be easily converted back to beta-carotenoids (Ruiz-sola et al., 2012).

The remaining candidate genes had functions that reduce the accumulation of carotenoids through carotenoid catabolism leading to biosynthesis of apocarotenoids such as absciscic acid. Candidate

genes *Manes.03G058000* and *Manes.03G057900* identified in chromosome 03 are *BETA-CAROTENE DIOXYGENASE* gene. The identified function of this gene was abscisic acid biosynthesis. Indeed, this function of *BETA-CAROTENE DIOXYGENASE* gene aligns with previous reports (Shin et al., 2022). Another candidate gene on chromosome 03 is *Manes.03G057750*, a *carotenoid cleavage dioxygenase (K11159*) gene which was identified to be involved in biosynthesis of abscisic acid. Previous reports confirm that this gene is involved in degrading carotenoids into apocarotenoids (Cheng et al., 2023). *Manes.04G055244* gene, a *Aldehyde oxidase / Retinal oxidase* gene identified on chromosome 04 was found to be involved in carotenoids catabolism in cassava roots. Earlier reports indicate that this gene is indeed involved in carotenoids catabolism in rice (Colasuonno et al., 2017). Candidate genes on chromosomes 14 and 18 could only be identified based on their biological process of apocarotenoids biosynthesis. The gene ontology analysis of these genes highlighted those on chromosome 14 as GO:0043288, GO:0043289 and those in chromosome 18 as GO:0043289, GO:0009688. It has been hypothesized that the negative correlation between dry matter content (DMC) and pVAC in cassava roots of African cassava germplasm could be due to presence of genes involved in catabolism of carotenoids to apocarotenoids (Ceballos et al., 2017).

The findings that *Manes.01G124200* and *Manes.09G079657* genes are involved in carotenoids biosynthesis activities suggest that breeders will have to select for these genes since they promote accumulation of pVAC in cassava roots. In the carotenoids biosynthesis pathway, beta carotene which is the predominant pVAC (Rodriguez-amaya & Kimura, 2004) is degraded into apocarotenoids such as abscisic acid. Therefore, selection against *Manes.03G058000*, *Manes.03G057900, Manes.03G057750, Manes.04G055244* genes may lead to development of novel cassava varieties with high level of pVAC.

Functional analysis of candidate genes arising from haplotype-based GWAS resulted to another set of genes involved in carotenoid anabolism and catabolism in cassava roots. Among these genes, was *Manes.01G001200* (Zeta-carotene isomerase / 15-cis-zeta-carotene isomerase), which is located in chromosome 01. This gene has been previously reported to be a candidate gene for accumulation of carotenoids in cassava roots (Bredeson et al., 2016; Ikeogu et al., 2019). *Manes.05G193700* that was detected in chromosome 05 is a 15-cis-phytoene desaturase (PDS, crtP) gene which is involved in carotenoid biosynthesis in plants by converting Phytoene to Zeta-carotene (Chen et al., 2023; Odipio et al., 2017). On chromosome 08 *Manes.08G037100* which is prolycopene isomerase (crtISO, crtH) gene was identified. Literature review showed that in plants, this gene is responsible in converting cis-carotenoids to all-trans carotenoids (Park et al., 2002) that have high vitamin A activity (Rodriguez-amaya & Kimura, 2004). Therefore, selection for these genes using genomic breeding tools such as MAS may enhance pVAC in cassava roots.

Other identified genes are involved in carotenoid catabolism, where they break carotenoids into smaller apocarotenoids consequently reducing the pVAC content in cassava roots. These genes included *Manes.03G083500* on chromosome 03 which is a 9-cis-epoxycarotenoid dioxygenase (NCED). This gene has been reported as pivotal in biosynthesis of abscisic acid to facilitate drought tolerance in plants (He et al., 2018). On the same chromosome *Manes.03G150400* which is a BETA-CAROTENE DIOXYGENASE gene was identified. This gene has been reports to be a major negative regulator of carotenoids in plants (Gonzalez-Jorge et al., 2013). On chromosome 05 there was *Manes.05G082900* (beta-carotene isomerase (DWARF27) gene) that is responsible for isomerizing all-trans-β-carotene to 9-cis-β-carotene, which is a crucial step in the strigolactone biosynthesis pathway in plants (Abuauf et al., 2018). On the same chromosome is *Manes.05G051700* (9-cis-beta-carotene 9',10'-cleaving dioxygenase gene) and *Manes.05G005000*

(Beta-carotene isomerase) both involved in biosynthesis of strigolactones. Other genes involved in carotenoid catabolism in cassava were *Manes.08G016300* which is a carotene epsilon-monooxygenase (LUT1, CYP97C1) gene, *Manes.10G141300* (CAROTENOID 9,10(9',10')-CLEAVAGE DIOXYGENASE 1), which is involved in apocarotenoid biosynthesis, *Manes.11G097600* which is a Phytoene desaturase (zeta-carotene-forming) / 2-step phytoene desaturase gene involved in zeaxanthin biosynthesis and *Manes.14G022700* (Beta-carotene 3-hydroxylase / Beta-carotene 3, 3'-monooxygenase), respectively, detected on chromosomes 08, 10, 11 and 14.

5.5. Conclusions

This study identified new genomic regions on chromosomes 03, 05, 06, 08, 09, 10, 11 and 18 that are strongly associated with pVAC. Haplotype-based GWAS identified 15 regions (haplotype blocks) on chromosome 01, 03, 04, 05, 06, 08, 10, 11, 14 and 18 significantly associated with the traits. This was the highest number of significant SNPs identified by a single model in this study. The use of multi-locus random marker effect GWAS identified five SNPs on chromosomes 03, 04, 09, 14 and 18 that were significantly associated with beta-carotene content in cassava roots whereas the traditional GWAS using BLINK model identified two significant SNPs on chromosome 09 and 14. The mrMLM model identified majority (4 out of 5) of significant SNPs compared to other models of multi-locus random marker effect GWAS. It was concluded that the use of haplotype-based GWAS model provides high statistical power to detect more genomic regions that are associated with pVAC in cassava.

The identified SNPs revealed significant allelic effects on beta carotene accumulation in cassava roots. Favorable alleles associated with higher beta-carotene content provide useful targets for

marker-assisted selection and genetic improvement. The fixation of the locus on chromosome 18 suggests prior selection pressure and potential stability of the trait in breeding populations. GWAS allele effects based on additive models can misrepresent biological reality. Genotype-phenotype plots reveal non-additive inheritance (dominance and epistasis) especially on chromosomes 03 and 14. Beta-carotene accumulation depends on specific allele combinations, emphasizing the need to analyze genotype-level effects rather than rely solely on GWAS averages.

A total of 20 candidate genes were found to be involved in carotenoid biosynthesis pathway in cassava. Among the identified candidate genes, *Manes.01G124200*, *Manes.01G001200*, *Manes.05G193700, Manes.08G037100*, and *Manes.03G084700* are involved in carotenoid accumulation and thus, cassava breeders may select in their favor. The remaining 15 genes were found to be involved in carotenoids catabolism and they were: *Manes.03G150400, Manes.03G083500, Manes.03G058000, Manes.03G057750, Manes.03G057900, Manes.04G055244, Manes.05G082900, Manes.05G051700, Manes.05G005000, Manes.08G016300, Manes.09G079657, Manes.09G144600, Manes.10G141300, Manes.11G097600* and Manes.*14G022700.*

CHAPTER SIX

6.0 GENERAL CONCLUSIONS AND RECOMMENDATIONS

6.1 General conclusions

Results showed that the SNP markers used in this study were reliable since they were well distributed across the genome and also having PIC mean value of 0.24 and MAF mean value of 0.20. Genetic diversity analysis showed that the whole population had moderately diverse alleles, which breeders can use, as a source of novel alleles for development of new cassava varieties with high pVAC. There is a need to direct more effort to expand the genetic base of pVAC cassava population from moderate to high

The study population was structured into three sub-populations whereby variability within sub-population was higher compared to that of between sub-populations. This suggests that the subpopulations can be a source for selection of superior progenitors. The presence of high variability within population compared to between populations suggest that conservation efforts should be directed toward locally adapted genotypes without considering differences between subpopulations

At an LD threshold of r^2= 0.2, the genome-wide LD in the study population decreased or dissociated at 613.072 kb. This finding demonstrated that high-resolution genome-wide investigations may be conducted using a small number of markers (1,256) and candidate genes identified reliably from the study population. Furthermore, presence of LD in the study population provides a basis for identification of haplotypes. Presence of high LD in the population and haplotyping not only supports genomic analyses but also application of pyramiding breeding.

The prediction ability of basic GWAS model was low and with overfitting in the case of parametric models. However, adding significant SNPs from GWAS greatly enhanced the prediction ability of BGLR genomic prediction models. The parametric GP models improved further as more significant SNPs were included into the model. A multi-locus random marker effect GWAS discovered more causal SNPs in the cassava genome compared to traditional GWAS. The resulting non-parametric ML models were highly predictive and thus reliable in forecasting genomic value since they account for both additive and non-additive gene effects. Feature selection and the usage of an ensemble of models dramatically improved the models' computation speed and prediction accuracy.

The application of multi-locus random marker effect GWAS offers a reliable model to detect additional genomic areas associated with pVAC in cassava. However, haplotype-based GWAS had a higher detection power compared to the studied GWAS models. The identification of more significant SNPs by GWAS assisted the identification of more superior or favorable alleles for use in genetic improvement of pVAC cassava. The use of haplotype based and multi-locus GWAS models led to detection of new causative variants on chromosomes 03, 05, 06, 08, 09, 10, 11 and 18 that are strongly associated with pVAC discovered through this study.

Favorable alleles were uncovered from SNP markers that were significantly associated with beta-carotene content in cassava roots. This is of great importance in the implementation of pyramid breeding. Although significant SNP markers may only explain a small portion of phenotypic variance, pyramiding all advantageous alleles from several SNP marker loci into a single genotype may have a considerable impact, culminating in the formation of an elite cultivar. To improve cassava biofortification, favorable alleles for pVAC should be appropriately integrated via marker-assisted selection.

Functional analysis of candidate genes uncovered 20 genes controlling carotenoid biosynthesis in cassava roots. This suggests that carotenoid content in cassava roots could be controlled by a nexus of genes. Among the 20 genes, *Manes.01G124200*, Manes.01G001200, *Manes.05G193700, Manes.08G037100*, and *Manes.03G084700* are involved in carotenoid accumulation and thus, cassava breeders may select in their favor. Conversely, cassava breeders may have to select against the remaining genes that were found to be involved carotenoid in catabolism

6.2 Recommendations

Basing of the finding that the population had a moderate diversity, this study recommends further expansion of the genetic base through hybridization of the new pVAC accession with local varieties as a strategy to enhance diversity. To enhance the effectiveness of genomic selection models, it is recommended to use significant SNPs from marker-trait associations. Haplotype-based GWAS and multi locus random marker effect GWAS models are recommended since they showed higher detection power. Selection of final clone can reliably be carried out using RF, XGBTree and KNN models or their ensemble. It is recommended to employ feature selection as a strategy of optimizing the prediction power of these models. Pyramiding of the identified favorable alleles was recommended as a strategy of effective biofortification of cassava with pVAC. Cassava biofortification programs should target the 20 genes identified in genetic improvement of the crop with pVAC.

REFERENCES

Aalborg, T., Sverrisdottir, E., Kristensen, H. T., & Nielsen, K. L. (2024). The effect of marker types and density on genomic prediction and GWAS of key performance traits in tetraploid potato. *Frontiers in Plant Science*, *15*(March), 1–17. https://doi.org/10.3389/fpls.2024.1340189

Abed, A., & Belzile, F. (2019). Comparing Single-SNP Multi-SNP and Haplotype-Based Approaches in Association Studies forMajor Traits in Barley. *Plant Genome*, *12*(3).

Abincha, W., Dzidzienyo, D. K., Wesonga, L. N., Mwimali, M., Ozimati, A., Kayondo, I. S., Ofori, K., Tongoona, P., & Kivuva, B. M. (2024). Enhancing Nutrition : A Review of Provitamin A Carotenoid Cassava Breeding Initiatives in East Africa. *Advances in Agriculture*, *2024*. https://doi.org/10.1155/aia/9937876

Abincha, W., Ikeogu, U. N., Kawuki, R., Egesi, C., Rabbi, I., Parkes, E., Kulakow, P., Edema, R., Gibson, P., & Owor, B. E. (2021). Portable spectroscopy calibration with inexpensive and simple sampling reference alternatives for dry matter and total carotenoid contents in cassava roots. *Applied Sciences (Switzerland)*, *11*(4), 1–11. https://doi.org/10.3390/app11041714

Abincha, W., Kawuki, R., Ikeogu, U., Obara, J., Hussin, Y., Egesi, C., Kulakow, P., Rabbi, I., Owor, B.-E., & Parkes, E. (2020). Comparison of Near-infrared Spectroscopy with other options for total carotenoids content phenotyping in fresh cassava roots. *Journal of Scientific Agriculture*, *4*, 55–60. https://doi.org/10.25081/jsa.2020.v4.6196

Abuauf, H., Haider, I., Jia, K. P., Ablazov, A., Mi, J., Blilou, I., & Al-Babili, S. (2018). The Arabidopsis DWARF27 gene encodes an all-trans-/9-cis-β-carotene isomerase and is

induced by auxin, abscisic acid and phosphate deficiency. In *Plant Science* (Vol. 277, pp. 33–42). https://doi.org/10.1016/j.plantsci.2018.06.024

Adeniji, T. ., Sanni, L. ., Barimalaa, I. ., & Hart, A. . (2010). Mineral composition of five improved varieties of cassava. *Nigerian Food Journal*, *25*(2). https://doi.org/10.4314/nifoj.v25i2.50839

Akinwale, M. G., Aladesanwa, R. D., Akinyele, B. O., Dixon, A. G. O., & Odiyi, A. C. (2010). Inheritance of β-carotene in cassava (Manihot esculenta crantza). *International Journal of Genetics and Molecular Biology*, *2*(10), 198–201.

Aloqalaa, D. A., Kowalski, D. R., Błazej, P., Wnetrzak, M., Mackiewicz, D., & Mackiewicz, P. (2019). The impact of the transversion/transition ratio on the optimal genetic code graph partition. *BIOINFORMATICS 2019 - 10th International Conference on Bioinformatics Models, Methods and Algorithms, Proceedings; Part of 12th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2019*, *April*, 55–65. https://doi.org/10.5220/0007381000550065

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, P., & Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. *Nature Protocols*, *5*(9), 1564–1573. https://doi.org/10.1038/nprot.2010.116.Data

Arango, J., Wu, F., & Welsch, R. (2010). Characterization of phytoene synthases from cassava and their involvement in abiotic stress-mediated responses. *Planta*, *232*(October), 1250–1262. https://doi.org/10.1007/s00425-010-1250-6

Augusto, A., & Alves, C. (2002). Cassava Botany and Physiology. In R. . Hikllocks, J. . Thresh, & A. . Bellotti (Eds.), *Cassava: Biology, Production and Utilization* (pp. 67–89).

Badejo, A. A. (2018). Elevated carotenoids in staple crops : The biosynthesis , challenges and measures for target delivery. *Journal of Genetic Engineering and Biotechnology*, *16*(2), 553–562. https://doi.org/10.1016/j.jgeb.2018.02.010

Bajgain, P., & Anderson, J. A. (2021). Multi-Allelic Haplotype-Based Association Analysis Identifies Genomic Regions Controlling Domestication Traits in Intermediate Wheatgrass. *Agriculture*, *11*(667). https://doi.org/https://doi.org/10.3390/agriculture11070667

Balagopalan, C. (2002). Cassava Utilization in Food , Feed and Industry. In R. . Hillocks, J. . Thresh, & A. . Bellotti (Eds.), *Cassava: Biology, Production and Utilization* (pp. 301–318). CAB International.

Barrett, J. C., Fry, B., Maller, J., & Daly, M. J. (2005). Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics*, *21*(2), 263–265. https://doi.org/10.1093/bioinformatics/bth457

Belalcazar, J., Dufour, D., Meike, A. S., Pizarro, M., Luna, J., Londoño, L., Nelson, M., Angélica M., Jaramillo Lizbeth, P., Luis A., B. L.-L., Fabrice, D., & Elise F., Talsma, and Hernán, C. (2016). High-Throughput Phenotyping and Improvements in Breeding Cassava for Increased Carotenoids in the Roots. *Crop Science*, *56*, 2916–2925.

Belete, G. T., Fenta, A. L., & Hussen, M. S. (2019). Xerophthalmia and Its Associated Factors among School-Age Children in Amba Giorgis Town, Northwest Ethiopia, 2018. *Journal of Ophthalmology*, *2019*. https://doi.org/10.1155/2019/5130904

Bhagwat, S., Gulati, D., Sachdeva, R., & Sankar, R. (2014). Food fortification as a complementary strategy for the elimination of micronutrient deficiencies: Case studies of large scale food fortification in two Indian States. *Asia Pacific Journal of Clinical Nutrition*,

23(March), S4–S11. https://doi.org/10.6133/apjcn.2014.23.s1.03

Bredeson, J. V., Lyons, J. B., Prochnik, S. E., Wu, G. A., Ha, C. M., Edsinger-Gonzales, E., Grimwood, J., Schmutz, J., Rabbi, I. Y., Egesi, C., Nauluvula, P., Lebot, V., Ndunguru, J., Mkamilo, G., Bart, R. S., Setter, T. L., Gleadow, R. M., Kulakow, P., Ferguson, M. E., … Rokhsar, D. S. (2016). Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nature Biotechnology*, *34*(5), 562–570. https://doi.org/10.1038/nbt.3535

Burns, A., Gleadow, R., Cliff, J., Zacarias, A., & Cavagnaro, T. (2010). Cassava: The Drought, War and Famine Crop in a Changing World. *Sustainability*, *2*, 3572–3607. https://doi.org/10.3390/su2113572

Carneiro, M., Ferrand, N., & Nachman, M. W. (2009). Recombination and speciation: Loci near centromeres are more differentiated than loci near telomeres between subspecies of the European rabbit (Oryctolagus cuniculus). *Genetics*, *181*(2), 593–606. https://doi.org/10.1534/genetics.108.096826

Ceballos, H., Davrieux, F., & Talsma, E. (2017). Carotenoids in Cassava Roots. In *Carotenoids* (Issue June). https://doi.org/10.5772/intechopen.68279

Ceballos, H., Iglesias, C. A., Pérez, J. C., & Dixon, A. G. O. (2004). Cassava breeding: Opportunities and challenges. *Plant Molecular Biology*, *56*(4), 503–516. https://doi.org/10.1007/s11103-004-5010-5

Ceballos, H., Kawuki, R. S., Gracen, V. E., Yencho, G. C., & Hershey, C. H. (2015). Conventional breeding, marker-assisted selection, genomic selection and inbreeding in clonally propagated crops: a case study for cassava. In *Theoretical and Applied Genetics*

(Vol. 128, Issue 9). https://doi.org/10.1007/s00122-015-2555-4

Ceballos, H., Morante, N., Sánchez, T., Ortiz, D., Aragón, I., Chávez, A. L., Pizarro, M., Calle, F., & Dufour, D. (2013). Rapid cycling recurrent selection for increased carotenoids content in cassava roots. *Crop Science*, *53*(6). https://doi.org/10.2135/cropsci2013.02.0123

Ceballos, H., Pérez, J. C., Barandica, O. J., Lenis, J. I., Morante, N., Calle, F., Pino, L., & Hershey, C. H. (2016). Cassava breeding I: The value of breeding value. *Frontiers in Plant Science*, *7*(AUG2016), 1–12. https://doi.org/10.3389/fpls.2016.01227

Ceballos, H., Rojanaridpiched, C., & Phumichai, C. (2020). Excellence in Cassava Breeding : Perspectives for the Future. *Hapress Crop Breeding Genetics and Genomics*, *2*(2), 1–31.

Ceballos, H., Sánchez, T., Chávez, A. L., Iglesias, C., Debouck, D., Mafla, G., & Tohme, J. (2006). Variation in crude protein content in cassava (Manihot esculenta Crantz) roots. *Journal of Food Composition and Analysis*, *19*(6–7), 589–593. https://doi.org/10.1016/j.jfca.2005.11.001

Ceballos, H., Sánchez, T., Morante, N., Fregene, M., & Mestres, C. (2007). Discovery of an amylose- free starch mutant in cassava (Manihot esculenta Crantz). *Journal of Agricultural and Food Chemistry*, *55*(18), 7469–7476.

Chavez, A. L., Ceballos, H., Rodriguez-Amaya, D. ., Perez, J. ., Sanchez, T., Calle, F., & Morante, N. (2008). Sampling Variation for Carotenoids and Dry Matter Contents in Cassava Roots. *Journal of Root Crops*, *34*(1), 43–49.

Chen, H., Liang, M.-H., Ye, Z.-W., Zhu, Y.-H., & Jian, J.-G. (2023). Engineering-the-β-carotene-metabolic-pathway-of-microalgae-dunaliella-to-confirm-its-carotenoid Synthesis

Pattern in Comparison To Bacteria and Plants Hao-Hong. *Microbiology Spectrum*, *11*(2).

Cheng, C., Yang, R., Yin, L., Zhang, J., Gao, L., Lu, R., Yang, Y., Wang, P., Mu, X., Zhang, S., Zhang, B., & Zhang, J. (2023). Characterization of Carotenoid Cleavage Oxygenase Genes in Cerasus humilis and Functional Analysis of ChCCD1. *Plants*, *12*(2114).

Colasuonno, P., Marcotuli, I., Lozito, M. L., & Simeone, R. (2017). Characterization of Aldehyde Oxidase (AO) Genes Involved in the Accumulation of Carotenoid Pigments in Wheat Grain. *Frontiers in Plant Science*, *8*(May), 1–11. https://doi.org/10.3389/fpls.2017.00863

Combs, G. F., & McClung, J. (2017). The Vitamins. In *Encyclopedia of Dietary Supplements*. https://doi.org/10.1081/E-EDS-120022052

Crupi, P., Faienza, M. F., Naeem, M. Y., Corbo, F., Clodoveo, M. L., & Muraglia, M. (2023). Overview of the Potential Beneficial Effects of Carotenoids on Consumer Health and Well-Being. *Antioxidants*, *12*(1069).

Das, J. K., Salam, R. A., Kumar, R., & Bhutta, Z. A. (2013). Micronutrient fortification of food and its impact on woman and child health: a systematic review. *Systematic Reviews*, *2*, 67. https://doi.org/10.1186/2046-4053-2-67

de Carvalho, R. R. B., Cortes, D. F. M., e Sousa, M. B., de Oliveira, L. A., & de Oliveira, E. J. (2022). Image-based phenotyping of cassava roots for diversity studies and carotenoids prediction. *PLoS ONE*, *17*(1), 1–25. https://doi.org/10.1371/journal.pone.0263326

Desta, Z. A., & Ortiz, R. (2014). Genomic selection: Genome-wide prediction in plant improvement. *Trends in Plant Science*, *19*(9), 592–601.

https://doi.org/10.1016/j.tplants.2014.05.006

Dickinson, A. (2014). The Benefits of Nutritional Supplements. In *Jornal Público* (Fourth). Council for Responsible Nutrition (CRN).

Dramadri, I. O., Nkalubo, S. T., Kramer, D. M., & Kelly, J. D. (2021). Genome-wide association analysis of drought adaptive traits in common bean. *Crop Science*, *January*, 1–22. https://doi.org/10.1002/csc2.20484

El-Sharkawy, M. A. (1993). Drought-tolerant Cassava for Africa, Asia, and Latin America. *BioScience*, *43*(7), 441–451. https://doi.org/10.2307/1311903

El Touati, Y., Slimane, J. Ben, & Saidani, T. (2024). Adaptive Method for Feature Selection in the Machine Learning Context. *Engineering, Technology and Applied Science Research*, *14*(3), 14295–14300. https://doi.org/10.48084/etasr.7401

Esuma, W., Kawuki, R. S., Herselman, L., & Labuschagne, M. T. (2016a). Diallel analysis of provitamin A carotenoid and dry matter content in cassava (Manihot esculenta Crantz). *Breeding Science*, *66*(4), 627–635. https://doi.org/10.1270/jsbbs.15159

Esuma, W., Kawuki, R. S., Herselman, L., & Labuschagne, M. T. (2016b). Stability and genotype by environment interaction of provitamin A carotenoid and dry matter content in cassava in Uganda. *Breeding Science*, *66*(3), 434–443. https://doi.org/10.1270/jsbbs.16004

Esuma, W., Ozimati, A., Kulakow, P., Gore, M. A., Wolfe, M. D., Nuwamanya, E., Egesi, C., & Kawuki, R. S. (2021). Effectiveness of genomic selection for improving provitamin A carotenoid content and associated traits in cassava. *G3 Genes|Genomes|Genetics*, *July*. https://doi.org/10.1093/g3journal/jkab160

Esuma, W., Rubaihayo, P., Pariyo, A., Kawuki, R., Wanjala, B., Nzuki, I., Harvey, J. J., & Baguma, Y. (2012). Genetic Diversity of Provitamin A Cassava in Uganda. *Journal of Plant Studies*, *1*(1). https://doi.org/10.5539/jps.v1n1p60

FAO. (2006). Food Security. In *FAO's Agriculture and Development Economics Division* (Vol. 43, Issue 2). https://doi.org/10.1016/j.jneb.2010.12.007

FAO. (2013). *Save and Grow: Cassava*.

FAO, AUC, ECA, & WFP. (2023). Africa Regional Overview of Food Security and Nutrition:Statistics and Trends. In *Accra, FAO*. https://www.fao.org/3/ca2127en/CA2127EN.pdf

FAO (2024). *FAOSTAT: Production: Crops and livestock products*. Retrieved February 5, 2024, from https://www.fao.org/faostat/en/#data/QCL

FAO, & WHO. (2001). Human Vitamin and Mineral Requirements. In *Human vitamin and mineral requirements*. ftp://ftp.fao.org/es/esn/nutrition/Vitrni/pdf/TOTAL.pdf

Foley, J. K., Michaux, K. D., Mudyahoto, B., Kyazike, L., Cherian, B., Kalejaiye, O., Ifeoma, O., Ilona, P., Reinberg, C., Mavindidze, D., & MD, E. B. (2021). Scaling Up Delivery of Biofortified Staple Food Crops Globally : Scaling Up Delivery of Biofortified Staple Food Crops Globally : Paths to Nourishing Millions. *Food and Nutrition Bulletin*, *February*, 1–17. https://doi.org/10.1177/0379572120982501

Fukuda, W., Guevara, C. L., Kawuki, R., & Ferguson, M. E. (2010). *Selected morphological and agronomic descriptors for the characterization of cassava*. 38.

Gabriel, S., Schaffner, S., Nguyen, H., Moore, J., Roy, J., Blumenstiel, B., Higgins, J., DeFelice,

M., Lochner, M., Faggart, A., Liu-Cordero, M., Rotimi, C., Adeemo, A., Cooper, R., Ward, R., Lander, E., Daly, M., & Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science*, *296*(June), 2225–2229. http://dx.doi.org/10.1126/science.1069424

Gacheru, P. K., Abong, G. O., Okoth, M. W., Lamuka, P. O., Shibairo, S. A., & Katama, C. M. (2015). Cyanogenic content, aflatoxin level and quality of dried cassava chips and flour sold in Nairobi and coastal regions of Kenya. *Current Research in Nutrition and Food Science*, *3*(3), 197–206. https://doi.org/10.12944/CRNFSJ.3.3.03

Gaj, T., Sirk, S. J., Shui, S., & Liu, J. (2016). Genome-Editing Technologies: Principles and Applications. *Cold Spring Harbor Perspective in Biology*, *8*, 1–20.

Gegios, A., Amthor, R., Maziya-dixon, B., Egesi, C., Mallowa, S., & Nungo, R. (2010). Children Consuming Cassava as a Staple Food are at Risk for Inadequate Zinc , Iron , and Vitamin A Intake. *Plant Foods Human Nutrition*, *65*, 64–70. https://doi.org/10.1007/s11130-010-0157-5

Gervason A, M., Ben O., O., Bibianne W., W., Edith W. T, W., & Jared M., O. (2017). Evaluation of Cyanide Levels in Two Cassava Varieties (*Mariwa* and *Nyakatanegi*) Grown in Bar-agulu, Siaya County, Kenya. *Journal of Food and Nutrition Research*, *5*(11), 817–823. https://doi.org/10.12691/jfnr-5-11-4

Githunguri, C., & Gatheru, M. (2017). *Situational Analysis of Cassava Production, Processing and Marketing in Kenya* (Issue January).

Githunguri, C., Lung'ahi, E. G., Kabugu, J., & Musili, R. (2015). Cassava farming transforming livelihoods among smallholder farmers in mutomo a Semi-arid District in Kenya. *Climate Change Management*, *October*, 225–233. https://doi.org/10.1007/978-3-319-13000-2_20

Gomez, M. A., Lin, Z. D., Moll, T., Chauhan, R. D., Hayden, L., Renninger, K., Beyene, G., Taylor, N. J., Carrington, J. C., Staskawicz, B. J., & Bart, R. S. (2019). Simultaneous CRISPR Cas9-mediated editing of cassava eIF4E isoforms nCBP-1 and nCBP-2 reduces cassava brown streak disease symptom severity and incidence. *Plant Biotechnology Journal*, *17*, 421–434.

Gonzalez-Jorge, S., Ha, S. H., Magallanes-Lundback, M., Gilliland, L. U., Zhou, A., Lipka, A. E., Nguyen, Y. N., Angelovici, R., Lin, H., Cepela, J., Little, H., Buell, C. R., Gore, M. A., & DellaPenna, D. (2013). Carotenoid cleavage dioxygenase4 is a negative regulator of β-carotene content in arabidopsis seeds. *Plant Cell*, *25*(12), 4812–4826. https://doi.org/10.1105/tpc.113.119677

Gutschker, S., Rüscher, D., Rabbi, I., Rosado-Souza, Laise Pommerrenig, Benjamin Pauly, M., Robertz, S., Doorn, A., Schlereth, Armin Neuhaus, Ekkehard Fernie, A., Reinert, Stephan Sonnewald, U., & Zierer, W. (2024). Carbon usage in yellow-fleshed Manihot esculenta storage roots shifts from starch biosynthesis to cell wall and raffinose biosynthesis via the myo-inositol pathway. *The Plant Journal*. https://doi.org/10.1111/tpj.16909

Hamazaki, K., & Iwata, H. (2020). Rainbow: Haplotype-based genome-wide association study using a novel SNP-set method. *PLoS Computational Biology*, *16*(2), 1–17. https://doi.org/10.1371/journal.pcbi.1007663

He, R., Zhuang, Y., Cai, Y., Agüero, C. B., Liu, S., Wu, J., Deng, S., Walker, M. A., Lu, J., & Zhang, Y. (2018). Overexpression of 9-cis-epoxycarotenoid dioxygenase cisgene in grapevine increases drought tolerance and results in pleiotropic effects. *Frontiers in Plant Science*, *9*(August), 1–16. https://doi.org/10.3389/fpls.2018.00970

Hillocks, R. J., Thresh, J. ., & Bellotti, A. . (2002). Cassava in Africa Origins of Cassava in Africa. In *Cassava: Biology, Production and Utilization* (pp. 41–54). http://ciat-library.ciat.cgiar.org/articulos_ciat/cabi_06ch3.pdf

Howard, R., Carriquiry, A. L., & Beavis, W. D. (2014). Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. *G3: Genes, Genomes, Genetics*, *4*(6), 1027–1046. https://doi.org/10.1534/g3.114.010298

Howeler, R. (2017). Does Cassava Cultivation Degrade or Improve the Soil ? *A New Future for Cassava in Asia: Its Use as Food, Feed and Fuel to Benefit the Poor*, *December*. file:///C:/Users/user/AppData/Local/Temp/2010-6205HowelerDoescassavaproductiondegradeorimprovethesoil.pdf

I. Udoh, L., U. Agogbua, J., R. Keyagha, E., & I. Nkanga, I. (2022). Carotenoids in Cassava (Manihot esculenta Crantz). In *Carotenoids-New Perspective and Application* (Issue June). https://doi.org/10.5772/intechopen.105210

Ikeogu, U. N., Akdemir, D., Wolfe, M. D., Okeke, U. G., Chinedozi, A., Jannink, J. L., & Egesi, C. N. (2019). Genetic Correlation, Genome-Wide Association and Genomic Prediction of Portable NIRS Predicted Carotenoids in Cassava Roots. *Frontiers in Plant Science*, *10*(December), 1–11. https://doi.org/10.3389/fpls.2019.01570

Ikeogu, U. N., Davrieux, F., Dufour, D., Ceballos, H., Egesi, C. N., & Jannink, J. L. (2017). Rapid analyses of dry matter content and carotenoids in fresh cassava roots using a portable visible and near infrared spectrometer (Vis/NIRS). *PLoS ONE*, *12*(12), 1–17. https://doi.org/10.1371/journal.pone.0188918

Imam, F., Musilek, P., & Reformat, M. Z. (2024). Parametric and Nonparametric Machine

Learning Techniques for Increasing Power System Reliability: A Review. *Information*, *15*(1), 1–23. https://doi.org/10.3390/info15010037

Institute of Medicine Food and Nutrition. (2001). *DIETARY REFERENCE INTAKES for Vitamin A, Vitamin K, Arsenic, Boron, Chromium, Copper, Iodine, Iron, Manganese, Molybdenum, Nickel, Silicon, Vanadium, and Zinc*. NATIONAL ACADEMY PRESS Washington, D.C.

Isendahl, C. (2011). The Domestication and Early Spread of Manioc (Manihot Esculenta Crantz): A Brief Synthesis. *Latin American Antiquity*, *22*(4), 452–468. https://doi.org/10.2307/23072569

Iwuagwu, O. (2012). The spread of cassava (manioc) in Igboland, south-east Nigeria: A reappraisal of the evidence. *Agricultural History Review*, *60*(1), 60–76.

Jannink, J. L., Lorenz, A. J., & Iwata, H. (2010). Genomic selection in plant breeding: From theory to practice. *Briefings in Functional Genomics and Proteomics*, *9*(2), 166–177. https://doi.org/10.1093/bfgp/elq001

Jaramillo, A. M., Londono, F. L., Orozco, J. C., Patino, G., Belalcazar, J., Davrieux, F., & Talsma, E. F. (2018). A comparison study of five different methods to measure carotenoids in biofortified yellow cassava (Manihot esculenta). *Plos*, *13*(12), 1–14.

Jennings, D. L., & Iglesias, C. (2002). Breeding for Crop Improvement. In *Cassava: Biology, Production and Utilization* (pp. 149–166).

Jhu, M., Ellison, E. E., & Sinha, N. R. (2023). *CRISPR gene editing to improve crop resistance to parasitic plants*. *October*, 1–11. https://doi.org/10.3389/fgeed.2023.1289416

Kamanda, I., Blay, E. T., Asante, I. K., Danquah, A., Ifie, B. E., Parkes, E., Kulakow, P., Rabbi,

I., Conteh, A., Kamara, J. S., Mensah, H. K., Whyte, J. B. A., & Sesay, S. (2020). Genetic diversity of provitamin-A cassava (Manihot esculenta Crantz) in Sierra Leone. *Genetic Resources and Crop Evolution*, *67*(5), 1193–1208. https://doi.org/10.1007/s10722-020-00905-8

Kamau, J., Melis, R., Laing, M., Derera, J., Shanahan, P., & Ngugi, E. (2010). Combining the yield ability and secondary traits of selected cassava genotypes in the semi-arid areas of Eastern Kenya. *Journal of Plant Breeding and Crop Science*, *2*(7), 181–191. http://www.academicjournals.org/jpbcs

Kamau, J., Melis, R., Laing, M., Derera, J., Shanahan, P., & Ngugi, E. C. K. (2011). Farmers ' participatory selection for early bulking cassava genotypes in semi-arid Eastern Kenya. *Journal of Plant Breeding and Crop Science*, *3*(3), 44–52.

Karimi, K., Farid, A. H., Sargolzaei, M., Myles, S., & Miar, Y. (2020). Linkage Disequilibrium, Effective Population Size and Genomic Inbreeding Rates in American Mink Using Genotyping-by-Sequencing Data. *Frontiers in Genetics*, *11*(March), 1–11. https://doi.org/10.3389/fgene.2020.00223

Kawuki, R. ., Pariyo, A., T, A., Nuwamanya, E., Ssemakula, G., Tumwesigye, S., Bua, A., Baguma, Y., Omongo, C., Alicai, T., & Orone, J. (2011). A breeding scheme for local adoption of cassava (Manihot esculenta Crantz). *Journal of Plant Breeding and Crop Science Vol.*, *3*(July), 120–130.

Kim, S. A., Cho, C. S., Kim, S. R., Bull, S. B., & Yoo, Y. J. (2018). A new haplotype block detection method for dense genome sequencing data based on interval graph modeling of clusters of highly correlated SNPs. *Bioinformatics*, *34*(3), 388–397.

https://doi.org/10.1093/bioinformatics/btx609

Labroo, M. R., & Rutkoski, J. E. (2022). New cycle, same old mistakes? Overlapping vs. discrete generations in long-term recurrent selection. *BMC Genomics*, *23*(1), 1–15. https://doi.org/10.1186/s12864-022-08929-3

Lai, K. L., Ng, J. Y., & Srinivasan, S. (2014). Xerophthalmia and keratomalacia secondary to diet-induced vitamin A deficiency in Scottish adults. *Canadian Journal of Ophthalmology*, *49*(1), 109–112. https://doi.org/10.1016/j.jcjo.2013.09.003

Lamichhaney, S. (2019). A comparison of the association between large haplotype blocks under selection and the presence / absence of inversions. *Ecology and Evolution*, *February*, 4888–4896. https://doi.org/10.1002/ece3.5094

Latif, S., & Müller, J. (2015). Potential of cassava leaves in human nutrition: A review. *Trends in Food Science and Technology*, *44*(2), 147–158. https://doi.org/10.1016/j.tifs.2015.04.006

Li, D., Xu, Z., Gu, R., Wang, P., Lyle, D., Xu, J., Zhang, H., & Wang, G. (2019). Enhancing genomic selection by fitting largeeffect SNPs as fixed effects and a genotypeby-environment effect using a maize BC1F3:4 population. *PLoS ONE*, *14*(10), 1–15. https://doi.org/10.1371/journal.pone.0223898

Lisker, A., Maurer, A., Schmutzer, T., Kazman, E., Cöster, H., Holzapfel, J., Ebmeyer, E., Alqudah, A. M., Sannemann, W., & Pillen, K. (2022). A Haplotype-Based GWAS Identified Trait-Improving QTL Alleles Controlling Agronomic Traits under Contrasting Nitrogen Fertilization Treatments in the MAGIC Wheat. *Plants*, *11*(3508).

Liu, F., Schmidt, R. H., Reif, J. C., & Jiang, Y. (2019). Selecting Closely-Linked SNPs Based on

Local Epistatic Effects for Haplotype Construction Improves Power of Association Mapping. *G3 Genes|Genomes|Genetics*, *9*(December), 4115–4126.

Liu, Q., Yang, F., Zhang, J., Liu, H., Rahman, S., Islam, S., Ma, W., & She, M. (2021). Application of crispr/cas9 in crop quality improvement. *International Journal of Molecular Sciences*, *22*(8), 1–16. https://doi.org/10.3390/ijms22084206

Liu, Y., Wang, D., He, F., Wang, J., Joshi, T., & Xu, D. (2019). Phenotype Prediction and Genome-Wide Association Study Using Deep Convolutional Neural Network of Soybean. *Frontiers in Genetics*, *10*(November 2017). https://doi.org/10.3389/fgene.2019.01091

Lopez Villar, J. (2015). *Tackling hidden hunger : putting diet diversification at the centre*.

Lu, Y., Shah, T., Hao, Z., Taba, S., Zhang, S., Gao, S., Liu, J., Cao, M., Wang, J., Prakash, A. B., Rong, T., & Xu, Y. (2011). Comparative SNP and Haplotype Analysis Reveals a Higher Genetic Diversity and Rapider LD Decay in Tropical than Temperate Germplasm in Maize. *PLoS ONE*, *6*(9). https://doi.org/10.1371/journal.pone.0024861

Luo, X., Tomlins, K. I., Carvalho, L. J. C. B., Li, K., & Chen, S. (2018). The analysis of candidate genes and loci involved with carotenoid metabolism in cassava (Manihot esculenta Crantz) using SLAF-seq. *Acta Physiologiae Plantarum*, *40*(4), 1–11. https://doi.org/10.1007/s11738-018-2634-7

Luo, Z., Szczepanek, A., & Abdel-Haleem, H. (2020). Genome-Wide Association Study (GWAS) analysis of camelina seedling germination under salt stress condition. *Agronomy*, *10*(9), 1–14. https://doi.org/10.3390/AGRONOMY10091444

Lyons, D. M., & Lauring, A. S. (2017). Evidence for the selective basis of transition-to-

transversion substitution bias in two RNA viruses. *Molecular Biology and Evolution*, *34*(12), 3205–3215. https://doi.org/10.1093/molbev/msx251

Makokha, A. ., & Tunje, T. . (2005). Potential for alleviating vitamin A deficiency in East Africa through cassava and sweet potato tubers. *Crop Science*, *7*, 643–646.

Mangal, V., Kumar, L., Kumar, S., Saxena, K., & Sood, S. (2024). Heliyon Triumphs of genomic-assisted breeding in crop improvement. *Heliyon*, *10*(15), e35513. https://doi.org/10.1016/j.heliyon.2024.e35513

Maroya, N. G., Kulakow, P., Parkes, E. Y., Schweigert, F. J.,Kulakow, P. A., Parkes, E. Y., Friedrich, S. K., Schweigert, F., Alamu, E., & Maziya-, B. (2012). *Quantification of total carotene content of yellow root cassava genotypes using iCheck TM Carotene in comparison with the standard spectrophotometer method Contact Quantification of total carotene content of yellow root cassava genotypes using iCheck TM C. June*. https://doi.org/10.13140/2.1.4534.4007

Martínez, M. E., Jacobs, E. T., Baron, J. A., Marshall, J. R., & Byers, T. (2012). Dietary supplements and cancer prevention: Balancing potential benefits against proven harms. *Journal of the National Cancer Institute*, *104*(10), 732–739. https://doi.org/10.1093/jnci/djs195

Mbanjo, E. G. N., Rabbi, I. Y., Ferguson, M. E., Kayondo, S. I., Eng, N. H., Tripathi, L., Kulakow, P., & Egesi, C. (2021). Technological Innovations for Improving Cassava Production in Sub-Saharan Africa. *Frontiers in Genetics*, *11*(January). https://doi.org/10.3389/fgene.2020.623736

Meira, D., Meier, C., Olivoto, T., Nardino, M., Klein, L. A., Moro, E. D., Fassini, F., Marchioro,

V. S., & Souza, V. Q. De. (2019). Estimates of genetic parameters between and within black oat populations. *Bragantia*, *78*(1), 43–51. https://doi.org/10.1590/1678-4499.2018116

Mezzalira, I., Costa, C. J., Vieira, E. A., de Freitas, J. F., Silva, M. S., Denke, M. L., & da Silva, K. N. (2013). Pre-germination treatments and storage of cassava seeds and their correlation with emergence of seedlings. *Journal of Seed Science*, *35*(1), 113–118. https://doi.org/10.1590/S2317-15372013000100016

MoALFC, M. of A. L. F. and C. (2020). *Kenya Agri-Nutrition Implementation Strategy 2020-2025*.

Momen, M., Mehrgardi, A. A., Sheikhi, A., Kranis, A., Tusell, L., Morota, G., Rosa, G. J. M., & Gianola, D. (2018). Predictive ability of genome-assisted statistical models under various forms of gene action. *Scientific Reports*, *8*(1), 1–11. https://doi.org/10.1038/s41598-018-30089-2

Monke, E. A., Argwings-Kodhek, G., Avillez, F., Mukumbu, M., Pagiola, S., Sellen, D., & Winter-Nelson, A. (2019). Agricultural Policy in Kenya. *Agricultural Policy in Kenya: Issues and Processes*, *March*, 20–22. https://doi.org/10.7591/9781501737442

Montagnac, J. A., Davis, C. R., & Tanumihardjo, S. A. (2009). Nutritional Value of Cassava for Use as a Staple Food and Recent Advances for Improvement. *Comprehensive Reviews in Food Science and Food Safety*, *8*(3), 181–194. https://doi.org/10.1111/j.1541-4337.2009.00077.x

Moore, W. S., & DeFillippis, V. R. (1997). The Window of Taxonomic Resolution for Phylogenies Based on Mitochondrial Cytochrome b. In *Avian Molecular Evolution and Systems* (pp. 83–119).

Morillo, Y., Sánchez, T., Morante, N., Chávez, A. L., Ana Cruz Morillo, C., Bolaños, A., & Ceballos, H. (2012). Preliminary study of inheritance of the carotenoids content in roots from cassava (Manihot esculenta Crantz) segregating populations. *Acta Agronómica*, *61*(3), 253–264.

Mota, L. F. M., Arikawa, L. M., Santos, S. W. B., Fernandes Júnior, G. A., Alves, A. A. C., Rosa, G. J. M., Mercadante, M. E. Z., Cyrillo, J. N. S. G., Carvalheiro, R., & Albuquerque, L. G. (2024). Benchmarking machine learning and parametric methods for genomic prediction of feed efficiency-related traits in Nellore cattle. *Scientific Reports*, *14*(1), 1–14. https://doi.org/10.1038/s41598-024-57234-4

Mukhopadhyay, T., & Bhattacharjee, S. (2016). Genetic Diversity: Its Importance and Measurements. In A. Mir & N. Bhat (Eds.), *Conserving Biological Diversity: Multiscaled Approach* (2016th ed., Issue April 2021, pp. 251–295). Research India Publications.

Mulu-Mutuku, M. W., Odero-Wanga, D. A., Ali-Olubandwa, A. M., Maling'a, J., & Nyakeyo, A. (2013). Commercialisation of Traditional Crops: Are Cassava Production and Utilisation Promotion Efforts Bearing Fruit in Kenya? *Journal of Sustainable Development*, *6*(7). https://doi.org/10.5539/jsd.v6n7p48

Musungayi, E. M., Ngugi, K., Muthomi, J. W., Were, V. W., Olubayo, F. M., Nzuve, F. M., & Yuga, M. E. (2018). Evaluation of Resistance of Cassava Half-Sib Progenies to Cassava Mosaic Disease and Their Agronomic Performances in Western Kenya. *Journal of Agricultural Science*, *10*(12), 78. https://doi.org/10.5539/jas.v10n12p78

Mutoni, C. K., Nzuve, F. M., Miano, D. W., Kivuva, B. M., Ferguson, M. E., Kipkoech, Y. H., Iita, T. A., Box, P. O., Service, I., Plant, K., Station, Q., & Box, P. O. (2021). Prevalence of

Cassava Brown Streak and Cassava Mosaic Diseases in Lamu County Kenya. *East African Agricultural and Forestry Journal*, *85*, 1–4.

National Institutes of Health. (2023). *Vitamin A and Carotenoids*. https://ods.od.nih.gov/factsheets/VitaminA-HealthProfessional/

NBA. (2021). *Press release: NBA Board Approves Environmental Release Application for GM Cassava* (Issue June 2020). https://www.biosafetykenya.go.ke/index.php?option=com_content&view=article&id=94&Itemid=124#:~:text=The National Biosafety Authority (NBA,Environmental Release of GM Cassava.

Neimsemman, A., Abush, T., Rabbi, I., Tewodros, M., & Abtew, W. (2024). DArTSNPbased genetic diversity analyses in cassava (Manihot esculenta) genotypes sourced from different regions revealed high level of diversity within population. *BioRxiv A- PrePrint*.

Ning, C., Xie, K., Huang, J., Di, Y., Wang, Y., Yang, A., Hu, J., Zhang, Q., Wang, D., & Fan, X. (2022). Marker density and statistical model designs to increase accuracy of genomic selection for wool traits in Angora rabbits. *Frontiers in Genetics*, *13*(September), 1–9. https://doi.org/10.3389/fgene.2022.968712

Njenga, P., Edema, R., & Kamau, J. (2010). Variability in carotenoid content among introduced yellow-fleshed cassava clones Résumé. *Second RUFORUM Biennial Meeting 20 - 24 September 2010, Entebbe, Uganda*, *September*, 139–142.

Njenga, P., Edema, R., & Kamau, J. (2014). Combining ability for beta-carotene and important quantitative traits in a cassava f1 population. *Journal of Plant Breeding and Crop Science*, *6*(2), 24–30. https://doi.org/10.5897/jpbcs12.069

Njenga, P., Edema, R., Kamau, J., & Abong', G. (2022). Agronomic Performance and Carotenoid Content of Kenyan Yellow-Fleshed Cassava Clones. *Journal of Agricultural Science*, *14*(5), 128. https://doi.org/10.5539/jas.v14n5p128

Njenga, P., Edema, R., Kamau, J., & Ooko, G. (2018). *Agronomic performance and carotenoid content of Kenyan yellow-fleshed cassava clones* (Vol. 17, Issue 17).

Nti, I. K., Nyarko-Boateng, O., & Aning, J. (2021). Performance of Machine Learning Algorithms with Different K Values in K-fold CrossValidation. *International Journal of Information Technology and Computer Science*, *13*(6), 61–71. https://doi.org/10.5815/ijitcs.2021.06.05

Ntui, V. O., Tripathi, J. N., Kariuki, S. M., & Tripathi, L. (2024). Cassava molecular genetics and genomics for enhanced resistance to diseases and pests. *Molecular Plant Pathology*, *25*(1), 1–14. https://doi.org/10.1111/mpp.13402

Nutrition International. (2024). *KENYA PROGRAMS*.

Odipio, J., Alicai, T., Ingelbrecht, I., Nusinow, D. A., Bart, R., & Taylor, N. J. (2017). Efficient CRISPR/cas9 genome editing of phytoene desaturase in cassava. *Frontiers in Plant Science*, *8*(October), 1–11. https://doi.org/10.3389/fpls.2017.01780

Ojulong, H. F., Labuschagne, M. T., Herselman, L., & Fregene, M. (2010). Yield traits as selection indices in seedling populations of cassava. *Crop Breeding and Applied Biotechnology*, *10*, 191–196.

Okechukwu, R. U., & Dixon, A. G. O. (2009). Performance of Improved Cassava Genotypes for Early Bulking, Disease Resistance, and Culinary Qualities in an Inland Valley Ecosystem.

Agronomy Journal, *101*(5), 1258–1265. https://doi.org/10.2134/agronj2008.0077

Olayide

, P., Alexandersson, E., Tzfadia, O., Lenman, M., Gisel, A., & Stavolone, L. (2023). Transcriptome and metabolome profiling identify factors potentially involved in pro-vitamin A accumulation in cassava landraces. *Plant Physiology and Biochemistry*, *199*, 107713. https://doi.org/10.1016/j.plaphy.2023.107713

Olayide, P., Large, A., Stridh, L., Rabbi, I., Baldermann, S., Stavolone, L., & Alexandersson, E. (2020). Gene Expression and Metabolite Profiling of Thirteen Nigerian Cassava Landraces to Elucidate Starch and Carotenoid Composition. *Agronomy*, *10*(424), 1–16. https://doi.org/10.3390/agronomy10030424

Olsen, K. M., & Schaal, B. A. (1999). Evidence on the origin of cassava: Phylogeography of Manihot esculenta. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(10), 5586–5591. https://doi.org/10.1073/pnas.96.10.5586

Ortiz, D., Sánchez, T., Morante, N., Ceballos, H., Pachón, H., Myriam, C., Chávez, A. L., & Escobar, A. F. (2011). *Sampling strategies for proper quantification of carotenoid content in cassava breeding*. *3*(January), 14–23.

Park, H., Kreunen, S. S., Cuttriss, A. J., DellaPenna, D., & Pogson, B. J. (2002). Identification of the carotenoid isomerase provides insight into carotenoid biosynthesis, prolamellar body formation, and photomorphogenesis. *Plant Cell*, *14*(2), 321–332. https://doi.org/10.1105/tpc.010302

Parkes, E., Aina, O., Kingsley, A., Iluebbey, P., Bakare, M., Agbona, A., Akpotuzor, P.,

Labuschagne, M., & Kulakow, P. (2020). Combining ability and genetic components of yield characteristics, dry matter content, and total carotenoids in provitamin a cassava f1 cross-progeny. *Agronomy*, *10*(12). https://doi.org/10.3390/agronomy10121850

Pérez, P., & de los Campos, G. (2014). BGLR : A Statistical Package for Whole Genome Regression and Prediction. *Genetics*, *198*(2), 483–495.

Piero, N. M., Joan, M. N., Richard, O. O., Jalemba, M. A., Omwoyo, O. R., & Chelule, C. R. (2015). Analytical & Bioanalytical Determination of Cyanogenic Compounds Content in Transgenic Acyanogenic Kenyan Cassava (Manihot esculenta Crantz) Genotypes : Linking Molecular Analysis to Biochemical Analysis. *Journal of Analytical and Bioanalytical Techniques*, *6*(5). https://doi.org/10.4172/2155-9872.1000264

Pushpalatha, R., & Gangadharan, B. (2020). Is Cassava (Manihot esculenta Crantz) a Climate "Smart" Crop? A Review in the Context of Bridging Future Food Demand Gap. *Tropical Plant Biology*, *13*(3), 201–211. https://doi.org/10.1007/s12042-020-09255-2

Rabbi, I. Y., Udoh, L. I., Wolfe, M., Parkes, E. Y., Gedil, M. A., Dixon, A., Ramu, P., Jannink, J., & Kulakow, P. (2017a). Genome-Wide Association Mapping of Correlated Traits in Cassava: Dry Matter and Total Carotenoid Content. *The Plant Genome*, *10*(3), 1–14. https://doi.org/10.3835/plantgenome2016.09.0094

Rabbi, I. Y., Udoh, L. I., Wolfe, M., Parkes, E. Y., Gedil, M. A., Dixon, A., Ramu, P., Jannink, J., & Kulakow, P. (2017b). Genome-Wide Association Mapping of Correlated Traits in Cassava: Dry Matter and Total Carotenoid Content. *The Plant Genome*, *10*(3), 1–14. https://doi.org/10.3835/plantgenome2016.09.0094

Ren, D., An, L., Li, B., Qiao, L., & Liu, W. (2021). Efficient weighting methods for genomic

best linear-unbiased prediction (BLUP) adapted to the genetic architectures of quantitative traits. *Heredity*, *126*(2), 320–334. https://doi.org/10.1038/s41437-020-00372-y

Rodriguez-amaya, D. B., & Kimura, M. (2004). *HarvestPlus Handbook for Carotenoid Analysis*.

Ruiz-sola, A., Águila, M., Ruiz-sola, M. Á., & Rodríguez-concepción, M. (2012). Carotenoid Biosynthesis in Arabidopsis : A Colorful Pathway Published By : The American Society of Plant Biologists Carotenoid Biosynthesis in Arabidopsis : A Colorful Pathway. *BioOne*, *2012*(10). https://doi.org/10.1199/tab.0158

Rutkoski, J. E., Poland, J. A., Singh, R. P., Huerta-Espino, J., Bhavani, S., Barbier, H., Rouse, M. N., Jannink, J., & Sorrells, M. E. (2014). Genomic Selection for Quantitative Adult Plant Stem Rust Resistance in Wheat. *The Plant Genome*, *7*(3), 1–10. https://doi.org/10.3835/plantgenome2014.02.0006

Saediman, H., Limi, M. A., Rosmawaty, Arimbawa, P., & Indarsyih, Y. (2016). Cassava consumption and food security status among cassava growing households in southeast sulawesi. *Pakistan Journal of Nutrition*, *15*(12), 1008–1016. https://doi.org/10.3923/pjn.2016.1008.1016

Saggafu, S. ., Saha, H. ., & Mwololo, J. (2019). FARMERS' PERCEPTION ON NEW CASSAVA VARIETIES GROWN IN COASTAL KENYA. *Global Scientific Journals*, *7*(7), 1–3.

Saltzman, A., Birol, E., Bouis, H. E., Boy, E., De Moura, F. F., Islam, Y., & Pfeiffer, W. H. (2013). Biofortification: Progress toward a more nourishing future. *Global Food Security*, *2*(1), 9–17. https://doi.org/10.1016/j.gfs.2012.12.003

Salvador, E. M., Steenkamp, V., & Mccrindle, C. (2014). Production , consumption and nutritional value of cassava (Manihot esculenta , Crantz) in Mozambique : An overview. *Journal of Agricultural Biotechnology and Sustainable Development*, *6*(3). https://doi.org/10.5897/JABSD2014.0224

Sánchez, T., Ceballos, H., Dufour, D., Ortiz, D., Morante, N., Calle, F., Zum Felde, T., Domínguez, M., & Davrieux, F. (2014). Prediction of carotenoids, cyanide and dry matter contents in fresh cassava root using NIRS and Hunter color techniques. *Food Chemistry*, *151*, 444–451. https://doi.org/10.1016/j.foodchem.2013.11.081

Sanchez, T., Chavez, A., Ceballos, H., Rodriguez-Amaya, D., Nestel, P., & Ishitami, M. (2006). Reduction or delay of post-harvest physiological deterioration in cassava roots with higher carotenoid content. *Journal of the Science of Food and Agriculture*, *86*, 634–639. https://doi.org/https://doi.org/10.1002/jsfa.2371

Sanchez, T., Salcedo, E., Ceballos, H., Mafla, G., Morante, N., Calle, F., Perez, J. C., Debouck, D., Jaramillo, G., & Moreno, I. X. (2009). Screening of Starch Quality Traits in Cassava (Manihot esculenta Crantz). *Starch-Journal*, *61*, 12–19. https://doi.org/10.1002/star.200800058

Sehgal, D., Mondal, S., Crespo-Herrera, L., Velu, G., Juliana, P., Huerta-Espino, J., Shrestha, S., Poland, J., Singh, R., & Dreisigacker, S. (2020). Haplotype-Based, Genome-Wide Association Study Reveals Stable Genomic Regions for Grain Yield in CIMMYT Spring Bread Wheat. *Frontiers in Genetics*, *11*(December), 1–13. https://doi.org/10.3389/fgene.2020.589490

Serrote, C. M. L., Reiniger, L. R. S., Silva, K. B., Rabaiolli, S. M. dos S., & Stefanel, C. M.

(2020). Determining the Polymorphism Information Content of a molecular marker. *Gene*, *726*(October 2019), 144175. https://doi.org/10.1016/j.gene.2019.144175

Sesay, J. V., Lebbie, A., Wadsworth, R., Nuwamanya, E., Bado, S., & Norman, P. E. (2023). Genetic Structure and Diversity Study of Cassava (Manihot esculenta) Germplasm for African Cassava Mosaic Disease and Fresh Storage Root Yield. *Open Journal of Genetics*, *13*, 23–47. https://doi.org/10.4236/ojgen.2023.131002

Sheng, Y., Zheng, W., Pei, K., & Ma, K. (2005). Genetic variation within and among populations of a dominant desert tree Haloxylon ammodendron (Amaranthaceae) in China. *Annals of Botany*, *96*(2), 245–252. https://doi.org/10.1093/aob/mci171

Shin, K., Seo, M., & Kim, Y. (2022). Molecular Properties of β -Carotene Oxygenases and Their Potential in Industrial Production of Vitamin A and Its Derivatives. *Antioxidants*, *11*(1180), 1–13.

Sibhatu, K. T., Krishna, V. V., & Qaim, M. (2015). Production diversity and dietary diversity in smallholder farm households. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(34), 10657–10662. https://doi.org/10.1073/pnas.1510982112

Singh, N. K., Joshi, A., Sahoo, S., Tufchi, M., & Rakshit, S. (2022). Molecular breeding for improving yield in maize: Recent advances and future perspectives. In *Qtl Mapping in Crop Improvement: Present Progress and Future Perspectives*. Elsevier Inc. https://doi.org/10.1016/B978-0-323-85243-2.00010-6

Sivabharathi, R. C., Rajagopalan, V. R., Suresh, R., Sudha, M., Karthikeyan, G., Jayakanthan, M., & Raveendran, M. (2024). Haplotype-based breeding: A new insight in crop improvement. *Plant Science*, *346*(March). https://doi.org/10.1016/j.plantsci.2024.112129

Smaragdov, M. G., & Kudinov, A. A. (2020). Assessing the power of principal components and wright's fixation index analyzes applied to reveal the genome-wide genetic differences between herds of Holstein cows. *BMC Genetics*, *21*(1), 1–15. https://doi.org/10.1186/s12863-020-00848-0

Soto, J. C., Ortiz, J. F., Perlaza-jiménez, L., Vásquez, A. X., Lopez-lavalle, L. A. B., Mathew, B., Léon, J., Bernal, A. J., Ballvora, A., & López, C. E. (2015). A genetic map of cassava (Manihot esculenta Crantz) with integrated physical mapping of immunity-related genes. *BMC Genomics*, 16–190. https://doi.org/10.1186/s12864-015-1397-4

Souza, R. de L., Dias, L. A. D. S., Corrêa, T. R., Caixeta, E. T., Fernandes, E. da C., Muniz, D. R., Rosmaninho, L. B. de C., & Cardoso, P. M. R. (2019). Genetic variability revealed by microsatellite markers in a germplasm collection of jatropha curcas l. In brazil: An important plant for biofuels. *Crop Breeding and Applied Biotechnology*, *19*(3), 337–346. https://doi.org/10.1590/1984-70332019v19n3a46

Srinivasan, R., Babu, S., & Gothandam, K. M. (2017). Accumulation of phytoene , a colorless carotenoid by inhibition of phytoene desaturase (PDS) gene in Dunaliella salina V-101. *Bioresource Technology*, *242*, 311–318. https://doi.org/10.1016/j.biortech.2017.03.042

Ssemakula, G., Dixon, A. G. O., & Maziya-Dixon, B. (2007). Stability of total carotenoid concentration and fresh yield of selected yellow-fleshed cassava (Manihot esculenta Crantz). *Journal of Tropical Agriculture*, *45*(1–2), 14–20.

Stanley, L., & Yuan, Y. (2019). Transcriptional Regulation of Carotenoid Biosynthesis in Plants : So Many Regulators , So Little Consensus. *Frontiers in Plant Science*, *10*(August), 1–17. https://doi.org/10.3389/fpls.2019.01017

Stephenson, K., Amthor, R., Maziya-Dixon, B., Mallowa, S., Nungo, R., Gichuki, S., Mbanaso, A., & Manary, M. J. (2010). Children consuming cassava as a staple food are at risk for inadequate zinc, iron, and vitamin A intake. *Plant Foods for Human Nutrition*, *65*(1), 64–70. https://doi.org/10.1007/s11130-010-0157-5

Tabangin, M. E., Woo, J. G., & Martin, L. J. (2009). The effect of minor allele frequency on the likelihood of obtaining false positives. *BMC Proceedings*, *4*, 5–8. https://doi.org/10.1186/1753-6561-3-S7-S41

Talsma, E. F., Brouwer, I. D., Verhoef, H., Mbera, G. N. K., Mwangi, A. M., Demir, A. Y., Maziya-Dixon, B., Boy, E., Zimmermann, M. B., & Melse-Boonstra, A. (2016). Biofortified yellow cassava and Vitamin A status of Kenyan children: A randomized controlled trial. *American Journal of Clinical Nutrition*, *103*(1), 258–267. https://doi.org/10.3945/ajcn.114.100164

Tanumihardjo, S. A., Palacios, N., & Pixley, K. V. (2010). Provitamin A Carotenoid Bioavailability : What Really Matters ? *Intational Journal for Vitamin and Nutritional Research*, *80*(February 2015), 336–350. https://doi.org/10.1024/0300-9831/a000042

Temesgen, Z., Basa, B., & Herago, T. (2019). Medicinal , Nutritional and Anti-Nutritional Properties of Cassava (Manihot esculenta): A Review. *Academic Journal of Nutrition*, *8*(3), 34–46. https://doi.org/10.5829/idosi.ajn.2019.34.46

Tonukari, N. J. (2004). Cassava and the future of starch. *Electronic Journal of Biotechnology*, *7*(1), 12–15. https://doi.org/10.2225/vol7-issue1-fulltext-9

Tsetsos, F., Drineas, P., & Paschou, P. (2018). Genetics and population analysis. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, *1–3*, 363–378.

https://doi.org/10.1016/B978-0-12-809633-8.20114-3

Udogu, O. ., Omosun, G., & Njoku, D. . (2021). Comparative Evaluation of Physiological Post-harvest Root Deterioration, Total Carotenoids, Starch Content and Dry Matter of Selected Cassava Cultivars. *Nigerian Agricultural Journal*, *52*(1), 219–226. http://academicjournals.org/article/article1381317405_Salcedo et al.pdf

Udoh, L., Agogbua, J. U., Keyagha, E. R., & Nkanga, I. I. (2022). Carotenoids in Cassava (Manihot esculenta Crantz). In *Carotenoids-New Perspective and Application*. https://doi.org/http://dx.doi.org/10.5772/intechopen.105210 took

UN. (2003). Poverty and hunger. *Bulletin on the Eradication of Poverty - Time to End Poverty*, *10*, 1–17.

UN, N. U. (2022). World Population Prospects 2022. In *United Nation* (Issue 9). www.un.org/development/ desa/pd/.

United Nations. (2019). World Population Prospects 2019. In *Department of Economic and Social Affairs. World Population Prospects 2019.* (Issue 141). http://www.ncbi.nlm.nih.gov/pubmed/12283219

USAID. (2017). *Vulnerability, impacts and adaptation assessment in the East Africa region.* (Issue October). https://www.climatelinks.org/sites/default/files/asset/document/2017_USAID-PREPARED-TetraTech_Vulnerability-Impacts-Adaptation-Assessment-East-Africa-Food-Security-Agriculture.pdf

Varona, L., Legarra, A., Toro, M. A., & Vitezica, Z. G. (2018). Non-additive effects in genomic

selection. *Frontiers in Genetics*, *9*(MAR), 1–12. https://doi.org/10.3389/fgene.2018.00078

Voss-Fels, K. P., Cooper, M., & Hayes, B. J. (2019). Accelerating crop genetic gains with genomic selection. *Theoretical and Applied Genetics*, *132*(3), 669–686. https://doi.org/10.1007/s00122-018-3270-8

Wambua, M., Mulwa, R. M. S., Arama, P. F., Atieno, S. A., & Ogendo, J. O. (2020). Evaluation of popular cassava varieties for yield and cyanide content under ASAL conditions in Kenya. *African Crop Science Journal*, *28*(s1), 71–82. https://doi.org/10.4314/acsj.v28i1.6s

Wang, J., & Zhang, Z. (2021). GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and Prediction. *Genomics, Proteomics and Bioinformatics*, *19*(4), 629–640. https://doi.org/10.1016/j.gpb.2021.08.005

Wang, N., Akey, J. M., Zhang, K., Chakraborty, R., & Jin, L. (2002). Reconstruction of magnetic field surges to the poles from sunspot impulses. *American Journal of Human Genetics.*, *71*(1227), 1234. https://doi.org/10.1017/S1743921312004681

Wang, S. B., Feng, J. Y., Ren, W. L., Huang, B., Zhou, L., Wen, Y. J., Zhang, J., Dunwell, J. M., Xu, S., & Zhang, Y. M. (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Scientific Reports*, *6*(May 2015), 1–10. https://doi.org/10.1038/srep19444

Wang, Y. H., Zou, C. Q., Mirza, Z., Li, H., Zhang, Z. Z., Li, D. P., Xu, C. L., Zhou, X. Bin, Shi, X. J., Xie, D. T., He, X. H., & Zhang, Y. Q. (2016). Cost of agronomic biofortification of wheat with zinc in China. *Agronomy for Sustainable Development*, *36*(3). https://doi.org/10.1007/s13593-016-0382-x

Weber, S. E., Frisch, M., Snowdon, R. J., & Voss-fels, K. P. (2023). Haplotype blocks for genomic prediction : a comparative evaluation in multiple crop datasets. *Frontiers in Plant Science*, *September*, 1–17. https://doi.org/10.3389/fpls.2023.1217589

Welsch, R., Arango, J., Bär, C., Salazar, B., Al-Babili, S., Beltrán, J., Chavarriaga, P., Ceballos, H., Tohme, J., & Beyera, P. (2010). Provitamin a accumulation in cassava (Manihot esculenta) roots driven by a single nucleotide polymorphism in a phytoene synthase gene. *Plant Cell*, *22*(10), 3348–3356. https://doi.org/10.1105/tpc.110.077560

Wenjie, L., Li, W., Song, Z., Gao, Z., Xie, K., Wang, Y., Wang, B., Hu, J., Zhang, Q., Ning, C., Wang, D., & Fan, X. (2024). Marker Density and Models to Improve the Accuracy of Genomic Selection for Growth and Slaughter Traits in Meat Rabbits. *Genes*, *15*(4). https://doi.org/10.3390/genes15040454

White, W. L. B., Arias-garzon, D. I., Mcmahon, J. M., & Sayre, R. T. (1998). Cyanogenesis in Cassava: The Role of Hydroxynitrile Lyase in Root Cyanide Production. *Plant Physiology*, *116*(1998), 1219–1225.

Wu, X., Id, W. J., Id, C. F., Huang, J., & Zhou, G. (2022). Prioritized candidate causal haplotype blocks in plant genome-wide association studies. *PLoS Genetics*, *18*(10), 1–25. https://doi.org/10.1371/journal.pgen.1010437

Yoosefzadeh-Najafabadi, M., Rajcan, I., & Eskandari, M. (2022). Optimizing genomic selection in soybean: An important improvement in agricultural genomics. *Heliyon*, *8*(11), e11873. https://doi.org/10.1016/j.heliyon.2022.e11873

Zaidi, S. S., Mukhtar, M. S., & Mansoor, S. (2018). Genome Editing : Targeting Susceptibility Genes for Plant Disease Resistance. *Trends in Biotechnology*, *36*(9), 898–906.

https://doi.org/10.1016/j.tibtech.2018.04.005

Zhang, C., Dong, S., Xu, J., He, W., & Yang, T. (2019). *Genetics and population analysis PopLDdecay : a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files*. *35*(October 2018), 1786–1788. https://doi.org/10.1093/bioinformatics/bty875

Zhang, Y., Jin, J., Wang, N., Sun, Q., Feng, D., Zhu, S., Wang, Z., Li, S., Ye, J., Chai, L., Xie, Z., & Deng, X. (2024). Cytochrome P450 CitCYP97B modulates carotenoid accumulation diversity by hydroxylating b -cryptoxanthin in Citrus. *Plant Communications*, *5*(6), 100847. https://doi.org/10.1016/j.xplc.2024.100847

Zhang, Y. M., Jia, Z., & Dunwell, J. M. (2019). Editorial: The applications of new multi-locus gwas methodologies in the genetic dissection of complex traits. *Frontiers in Plant Science*, *10*(February), 1–6. https://doi.org/10.3389/fpls.2019.00100

Zhang, Y. W., Tamba, C. L., Wen, Y. J., Li, P., Ren, W. L., Ni, Y. L., Gao, J., & Zhang, Y. M. (2020). mrMLM v4.0.2: An R Platform for Multi-locus Genome-wide Association Studies. *Genomics, Proteomics and Bioinformatics*, *18*(4), 481–487. https://doi.org/10.1016/j.gpb.2020.06.006

Zhao, Y., Pu, Y., Liang, B., Bai, T., Liu, Y., Jian, L., & Ma, Y. (2022). A study using single-locus and multi-locus genome-wide association study to identify genes associated with teat number in Hu sheep. *Animal Genetics*, *53*, 203–211.

Zhou, G., Zhu, Q., Mao, Y., Chen, G., Xue, L., Lu, H., Shi, M., Zhang, Z., Song, X., Zhang, H., & Hao, D. (2021). Multi-Locus Genome-Wide Association Study and Genomic Selection of Kernel Moisture Content at the Harvest Stage in Maize. *Frontiers in Plant Science*,

12(July), 1–13. https://doi.org/10.3389/fpls.2021.697688

Zou, C., Du, Y., Rashid, A., Ram, H., Savasli, E., Pieterse, P. J., Ortiz-Monasterio, I., Yazici, A., Kaur, C., Mahmood, K., Singh, S., Le Roux, M. R., Kuang, W., Onder, O., Kalayci, M., & Cakmak, I. (2019). Simultaneous Biofortification of Wheat with Zinc, Iodine, Selenium, and Iron through Foliar Treatment of a Micronutrient Cocktail in Six Countries [Research-article]. *Journal of Agricultural and Food Chemistry*, *67*(29), 8096–8106. https://doi.org/10.1021/acs.jafc.9b01829

Appendices

Appendix 1: Functional analysis of the candidate gene for pro-vitamin A carotenoids in cassava

HB	Chr.	Pos.	Gene ID	Gene Name	Function	GO ID
57	1	5443919	Manes.01G124200	phytoene synthase (crtB)	carotenoid biosynthesis	GO:0009058, GO:0016740
				Zeta-carotene isomerase / 15-cis-zeta-carotene isomerase	lycopene biosynthesis II (plants)	
			Manes.01G001200	Zeta-carotene isomerase / 15-cis-zeta-carotene isomerase	carotenoid biosynthesis	GO:0006835, GO:0016020, GO:0017153
					lycopene biosynthesis II (plants)	
256	3	30374336	Manes.03G084700	phytoene synthase (crtB)	carotenoid biosynthesis	GO:0009058, GO:0016740
					lycopene biosynthesis II (plants)	
			Manes.03G083500	9-cis-epoxycarotenoid dioxygenase (NCED)	abscisic acid biosynthesis	
			Manes.03G057900	BETA-CAROTENE DIOXYGENASE	abscisic acid biosynthesis	
			Manes.03G057750	carotenoid cleavage dioxygenase (K11159)	abscisic acid biosynthesis	
			Manes.03G058000	BETA-CAROTENE DIOXYGENAS	abscisic acid biosynthesis	
244	3	3220936	Manes.03G084700	phytoene synthase (crtB)	carotenoid biosynthesis	
					lycopene biosynthesis II (plants)	

			Manes.03G150400	BETA-CAROTENE DIOXYGENASE	abscisic acid biosynthesis	
			Manes.03G058000	BETA-CAROTENE DIOXYGENASE	abscisic acid biosynthesis	
			Manes.03G057900	BETA-CAROTENE DIOXYGENASE	abscisic acid biosynthesis	
523	5	790785	Manes.05G193700	15-cis-phytoene desaturase (PDS, crtP)	carotenoid biosynthesis	GO:0016117, GO:0016705, GO:0055114
					lycopene biosynthesis II (plants)	
			Manes.05G082900	beta-carotene isomerase (DWARF27)	5-deoxystrigol biosynthesis	
			Manes.05G051700	9-cis-beta-carotene 9',10'-cleaving dioxygenase	5-deoxystrigol biosynthesis	
			Manes.05G005000	Beta-carotene isomerase	biosynthesis of strigolactones	
718	8	35791856	Manes.08G037100	prolycopene isomerase (crtISO, crtH)	carotenoid biosynthesis	GO:0016117, GO:0016491, GO:0016853, GO:0055114
			Manes.08G016300	carotene epsilon-monooxygenase (LUT1, CYP97C1)	apocarotenoid biosynthesis	GO:0004497, GO:0005506, GO:0016705, GO:0020037, GO:0055114
					lutein biosynthesis	
887	10	1538515	Manes.10G141300	CAROTENOID 9,10(9',10')-CLEAVAGE DIOXYGENASE 1	Apocarotenoid biosynthesis	
916	11	27900387	Manes.11G097600	Phytoene desaturase (zeta-carotene-forming) / 2-step phytoene desaturase	zeaxanthin biosynthesis	GO:0016491, GO:0055114
1159	14	28153367	Manes.14G022700	Beta-carotene 3-hydroxylase / Beta-carotene 3,3'-monooxygenase	apocarotenoid biosynthesis	GO:0005506, GO:0006633,

						GO:0016491, GO:0055114