APPLICATIONS OF MULTITRAIT AND MULTI-KERNEL MODELS FOR GENOMIC SELECTION IN AFRICAN CASSAVA.

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Uche Godfrey Okeke

December 2017.

APPLICATIONS OF MULTITRAIT AND MULTI-KERNEL MODELS FOR GENOMIC SELECTION IN AFRICAN CASSAVA.

Uche Godfrey Okeke, Ph.D.

Cornell University 2017.

Genomic selection (GS) could help accelerate African cassava breeding towards the development of high yielding, high dry matter (DM), disease resistant and provitamin A varieties. This work addresses some issues for implementing GS in cassava. First, we evaluated multivariate and univariate GS models via prediction accuracies. Second, the genetic basis for DM content was investigated using the Regional Heritability Mapping (RHM) procedure. Lastly, the genetic basis for co-inheritance of DM, root color and fresh yield (FYLD) were investigated using the Regional co-heritability Mapping (RHM) procedure. Key lessons were: (1) Multitrait (MT) models for single location data offered 40% higher average prediction accuracies for genomic breeding values (GEBVs) of six target traits across 3 locations compared to single-trait (uT) models. (2) Multivariate multi-environment (ME) models also offered 12% higher average prediction accuracies compared to a compound symmetric multi-environment model (uE) parameterized as a univariate multi-kernel model for multi-year multi-environment data. (3) The RHM analysis identified segments associated with DM in white cassava on chromosomes 1, 4, 5, 10, 17,18 and on yellow cassava chromosome 1. Candidates extracted from genes adjacent to the RHM significant segments include: glycosyltransferases, serine-threonine kinases (SnRKs), invertases and fructose bisphosphate aldolase. Prediction accuracies from these candidates and all genes in the RHM significant regions sug-

gest that they may be tagging regions associated with DM. (4) Genome-wide segment correlations from the RcHM analysis in yellow cassava showed a limited prospect for high DM yellow cassava development but good prospects for high DM, high yielding white cassava development.

BIOGRAPHICAL SKETCH

Uche Godfrey Okeke hails from Awgbu in Orumba North LGA of Anambra State Nigeria but was born in Uruakpan, now Abia state Nigeria on June 3, 1985 to parents Alfred Chibuogwu Okeke and Joy Njideka Okeke. Uche Godfrey Okeke did his nursery and kindergarten education at TRICOMA schools in Aba, Nigeria. Thereafter he completed his primary education at the Christian Progressive primary school and Living Word Academy all in Aba. He graduated senior secondary school in July 2001 at the Living Word Academy secondary school. In Febuary 2006, Uche graduated with second class upper honors from the Department of Animal Sciences and Technology, Federal University of Technology Owerri, Nigeria majoring in animal breeding and genetics. His drive and passion for breeding and statistical genetics led him to complete a masters education at the university of Helsinki, Finland in 2012 majoring in Bioinformatics and Systems Biology. Before finishing his masters education, he was already working at the DNA sequencing and genomics lab, Institute of Biotechnology, Viiki, Helsinki as a bioinformatician. Uche joined the NextGen cassava breeding project in August 2013 as a PhD student at the Cornell University plant breeding and genetics department. As part of this project he has worked with scientists at the International Institute for Tropical Agriculture (IITA) including Dr. Peter Kulakow and Dr. Ismail Rabbi. Uche has a passion for quantitative genetics and the improvement of crops and livestock. He has been working on the development of linear mixed models for applications in cassava breeding. Uche looks forward to contributing his quota to the knowledge and science of quantitative genetics and breeding.

To Chukwu Okike for making all things possible and for freely giving me the wherewithal for every accomplishment in life.

Also to my lovely parents, Alfred Chibuogwu Okeke and Joy Njideka Okeke. My siblings Ifunanya, Amy and NZ. You have all provided me with all the love, care, prayers, advises and support that I needed to finish this task.

ACKNOWLEDGEMENTS

I would like to acknowledge all those who contributed to this work and my development as a scientist. I would like to give special thanks to my committee; Dr. Jean-Luc Jannink for giving me an opportunity in his lab and for providing me with vital advises and supervision needed to complete my PhD; Professor Susan McCouch for her very vital contributions, support, encouragement, guidance and advise through all the complexities in my PhD program; Professor Jason Mezey for his immensely important contributions, encouragement, guidance for my PhD program and especially for his teaching in his wonderful quantitative genomics course. I give special thanks also to Dr. Peter Kulakow, Dr. Ismail Rabbi, Ikpan Smith, Agbona Afolabi, Moshood Bakare, Idhigu Cynthia, Teddy Hanmakyugh and all the IITA crew for supervising or collecting all the data needed for my research. I will never forget Dr. Deniz Akdemir who mentored me through the process of understanding and computation of mixed models. I also thank Dr. Marnin Wolfe, Ugochukwu Ikeogu, Roberto Lozano, Ariel Chan, Alfred Ozimati, Olumide Alabi who contributed to my research in many ways. I thank Nicholas Santantonio and Itaraju Brum for stimulating my reasoning and for fruitful discussions on the 4th floor Bradfield white board. I also thank all members of the Cornell plant breeding community from whom I have learned so much. I thank Adrianne Brown and family who I met here in the USA for their support, love and kindness. I would also like to give special thanks and gratitude to all who funded my PhD research and education, The Bill and Melinda Gates Foundation and UKaid (Grant 1048542; http://www.gatesfoundation.org) and the NextGen Cassava breeding project. Lastly, I thank my family for their prayers and unconditional support always.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Rationale and significance:

Cassava (Manihot esculenta Crantz) ranks as the sixth most important staple crop consumed mostly in Africa, South America and Asia by over 500 million people [1]. It is an outbreeding species cultivated clonally using stem cuttings of about 15 -30 cm long [1-2]. Cassava is hardy and can give substantial yields in marginal land or low input systems [3-4]. It is mainly cultivated for its root consisting of water and dry matter (DM) [6-8]. Cassava DM is made up of 90% starch making it very attractive as a high calorie staple or other starch-dependent industries [6-8]. The high demand for cassava [9-10] has necessitated urgent and rapid genetic improvement of this crop [11-12]. Another related component is the goal of biofortification of this crop geared towards fortifying cassava roots with beta-carotene, a provitamin A precursor [13-14]. The target is towards health benefits brought about by consumption of provitamin A cassava especially critical for maternal and child health development [13-14]. Plant breeding techniques and especially genomic selection (GS) offer vital tools for meeting these improved productivity and biofortification targets for cassava [13-14].

Efforts have been made over several decades towards genetic improvement of cassava [11,16,17]. However with the recent developments in genomics, rapid gains can be achieved towards improved productivity, disease resistance and biofortification of cassava via GS [18-20]. GS can accelerate breeding for quantitative traits [19-20]. This is especially useful for African cassava with previously

low research investments [21] and incomplete pedigree records. Another useful component is the need to understand the genetic basis of some complex traits in cassava. This understanding will help in the development of tools relevant for accelerating gains in cassava. Lastly, so much is yet to be unraveled on the implementation process for GS in African cassava genetic improvement.

1.2 Objectives:

This work addresses four major issues relevant for the implementation of GS in the genetic improvement of African cassava. Again, the primary focus is for improved productivity and development of provitamin A cassava. These include:

1. Understanding the accuracies of univariate and multivariate genomic prediction models in African cassava.
2. Understanding the genetic basis for the inheritance of DM content in African white and yellow cassava subpopulations via Regional heritability Mapping (RHM).
3. Understanding the genetic basis for the co-inheritance of DM, root color and fresh root yield (FYLD) in different cassava subpopulations via the Regional co-heritability Mapping (RcHM) procedure.

1.3 Literature review:

1.3.1 Cassava genomics:

The cassava genome spans 770 megabases (Mb) in 18 chromosomes [18, 22]. The version 4.1 cassava genome spans 532.5 Mb and consists of 30,666 protein coding genes with 3,485 alternative transcripts, median exon and intron lengths of 148 and 166 respectively [18]. The average number of genes per chromosome is 1700. Repetitive sequences covered 37.5% of the genome [18] consisting mainly of long interspersed nuclear elements and long-terminal repeat elements [23]. The cassava genome contained 147 regulatory microRNAs with a good number of other non-coding RNAs [23]. The number of genes in the cassava genome represent a less gene dense genome compared to rice and soybean [18]. The level of linkage disequilibrium in cassava varies by chromosomes (Figure1.1, non-published data from Roberto Lozano) and especially for chromosomes 1 and 4 with introgressions [80,66].

1.3.2 African cassava genetic improvement:

The International Institute of Tropical Agriculture (IITA) is at the center of cassava breeding in sub-Saharan Africa [17,24]. Over the past 3 to 4 decades, IITA in partnership with national agricultural research programs (NARPs) in over 20 sub-Saharan African countries have been involved in the development and dissemination of improved cassava varieties [17,24]. A total of 206 improved cassava varieties were released by NARPs in partnership with IITA in 20 African countries between 1970 and 1998 [17]. This number has increased dramatically

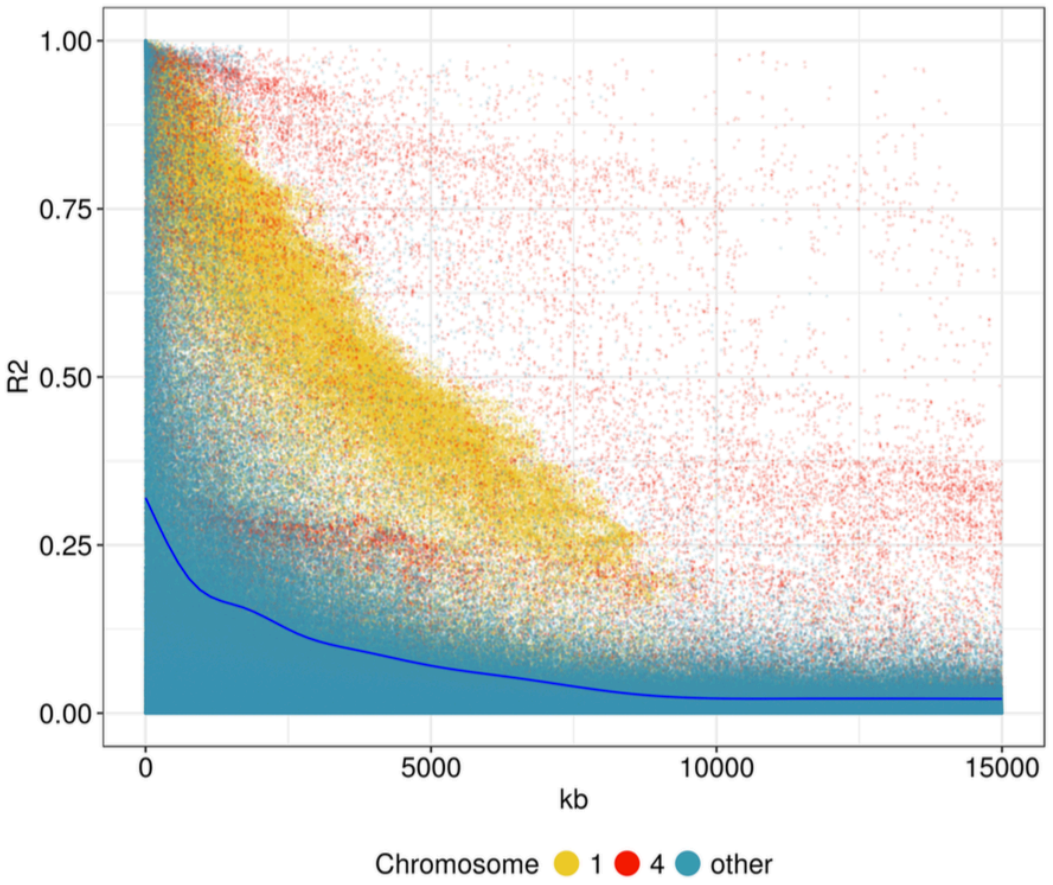

Figure 1.1: **Linkage disequilibrium decay in cassava.** LD decay calculated by pairwise correlation of GBS SNPs. Each dot represents a correlation value between SNPs at a chromosome and the blue line represents a smoothening line from a loess fit.

in recent times [17]. The majority of germplasm used in the development of these varieties were sourced from IITA breeding materials [17,24]. However, the germplasm base used to develop these breeding materials were assembled from local varieties in Africa, IITAs early breeding population, exotic materials from the International Center for Tropical Agriculture (CIAT) and wild cassava from Brazil [17,24].

In the 1970s, cassava genetic improvement was focused mainly on developing varieties resistant to cassava mosaic disease (CMD) and cassava bacterial blight (CBB) [17, 24-27]. Two geminiviruses including the African cassava mosaic virus (ACMV) and the East African cassava mosaic virus (EACMV) were responsible for the CMD [17, 24-26]. These viruses are transmitted by whitefly (*Bemisia tabaci*) and resulted in economic losses estimated then at 2 billion USD annually [17]. CBB was first reported in 1972 in Nigeria and was caused by *Xanthamonas campestris pv. manihotis* [17, 25]. IITA initiated breeding for resistance against CMD and CBB in 1971 using a resistant genotype (No. 58308) developed by the pioneer cassava breeding program in Africa at the Amani Research Station, Tanzania in 1930s. [24, 27-28]. Within a decade from 1971, several CMD and CBB resistant clones were developed from this program [17]. These elite clones also had stable yields and were accepted by consumers [17]. In summary, the breeding goal of the cassava breeding program of the IITA between 1970 and 1980 were resistance to CMD and CBB, consumer acceptability and yield stability [17, 27]. Germplasm from this early breeding effort were referred to as IITA early breeding population. This stock were founders and members of the Genetic Gain population. The success of this program was seen in a 1998 IITA survey which showed that improved varieties released by NARPs and IITA covered 22% of the 9 million hectares of cassava planted in 20 different sub-Saharan African countries [17]. These varieties accounted for 49% yield increase compared to the average yield and an annual increase of 10 million tons of fresh root [17]. Planting these improved varieties resulted in a gain of 204 USD per hectare [17].

However at the onset of the 1990s, a multi-disciplinary approach termed the Collaborative Study of Cassava in Africa (COSCA) was used to generate a

wealth of user information from local stakeholders in cassava including farmers, processors and marketers [29-31]. The COSCA study unraveled the attributes desired by farmers in six cassava producing countries (Congo, Cote d'ivoire, Ghana, Nigeria, Tanzania, and Uganda) and found that farmers preference for varieties were based on: high yield, earliness of bulking (early maturing), weed suppression ability, desirable branching habits, in-ground storability, pest and disease tolerance, low cyanogen level of storage roots, ease of peeling, mealiness after cooking, drought tolerance, high leaf yield, and ease of harvesting [29-30]. This information was used to re-organize the cassava breeding program at IITA in the 1990s leading to emphasis on target traits including: seedling vigor (VIGOR), Number of storage roots per hectare at harvest (RTNO), Fresh weight of harvested roots expressed in tons per hectare (T/ha) (FYLD), Harvest index (HI) measured as ratio of root weight to total biomass, percent dry matter (DM) of storage roots, which measures root dry weight as the percentage of the root fresh weight, plot mean cassava mosaic disease severity (MCMDS), rated on a scale from 1 (no symptoms) to 5 (extremely severe), and plot mean cassava green mite (MCGM) severity, rated on a scale from 1 (no symptoms) to 5 (extremely severe). The cassava green mite is *Mononychellus tanajoa* [32] High throughput measurements of mealiness and drought tolerance are still challenging (Personal communication, Ismail Rabbi). These breeding efforts post 1990s have led to the development and release of a few more varieties combining some of these traits especially MCMDS, DM and FYLD [17, 33].

Breeding for high provitamin A levels in African cassava was initiated by the HarvestPlus initiative Discovery Phase I (2003 - 2008) [34-35]. The breeding goals for this program were targeted towards high yielding (FYLD), virus resistant (MCMDS) and at least 25% of the daily required Vitamin A for women and

children. The latter goal was hinged on the measurement of total carotene content in cassava roots (TCC) using a portable iCheck device [35], a near-infrared spectroscopy device (NIRS) and a visual scoring based on the degree of yellowness of the roots (PLPCOL) scored as 1 for white to cream and 2 for deep cream to deep yellow roots [35]. The HarvestPlus initiative Development Phase II (2009 - 2013) program was used to develop protocols for rapid screening of cassava clones and rapid recurrent selection schemes for improving TCC [35]. In the Democratic Republic of Congo (DRC), HarvestPlus has officially released a variety (I011661) in 2008 with 7 ppm provitamin A content amounting to 46% of the 2003 set goal of 15 ppm provitamin A in the yellow cassava storage root [35]. Breeding effort for provitamin A cassava now contributes as much as 50% of IITAs total cassava breeding effort [35].

1.3.3 African cassava genetic improvement: Selection scheme

Most early stages of genotype evaluation for the IITA cassava breeding program are conducted at Ibadan (7.40 N, 3.90 E), Nigeria. Ibadan is a typical southern rainforest region with a good distribution of rainfall annually. Other stations where both early and advanced cassava evaluation trials are carried out include: Abuja (9.06 N, 7.40 E), Akure (7.26 N, 5.19 E), Ikenne (6.88 N, 3.70 E), Ilorin (8.48 N, 4.55 E), Mokwa (9.30 N, 5.0 E), Ubiaja (6.66 N, 6.38 E), Onne (4.74 N, 7.04 E), Warri (5.56 N, 5.79 E), Zaria (11.30 N, 7.69 E), Akwa-Ibom (5.07 N, 7.89 E), Benue (7.58 N, 8.69 E), Calabar (4.98 N, 8.34 E), Imo (5.52 N, 7.11 E), Taraba (8.71 N, 10.97 E), Umudike (5.47 N, 7.54 E) and Anambra (6.19 N, 7.11 E). These represent a good percentage of the cassava production areas in Nigeria and also reflect a range of the agro-ecological zones that may be typical in the

sub-Saharan African belt. Cassava breeding starts by hybridization of selected clones. This is usually carried out at Ubiaja (6.66 N, 6.38 E) where breeders have observed that most clones flower profusely (Personal communication, Ikpan Smith). Subsequently, selected F1 botanical seeds resulting from this hybridization go through the following cycle:

1. **Seedling nursery (SN):** Selected F1 seeds are sown in about 8cm depth plastic containers or bags by families (usually half sib or full sib). After about 45 days post germination [36], these seedlings are transplanted to a ridged field at a $1m^2$ spacing. Here, data on CMD severity are collected at 3, 6 and 9 months after planting (MAP). At harvest, data on number of roots, root weight, PLPCOL and TCC are also collected. Based on these data, some seedlings are selected and then cloned in the next stage (Personal communication, Ikpan Smith). It is worth mentioning that roots from cassava seedlings are fibrous thus inedible. This is why measurement of critical traits like DM and FYLD are skipped at this stage.
2. **Clonal evaluation trials (CETs):** The CET is a very critical stage in cassava breeding. It is the stage at which selection is imposed based on the breeding goals set using information from COSCA. The design of the CET is augmented with about 10-25 blocks depending on the size and area of the field. Each block contains about 15 - 20 clone accessions with 2 known checks in each block. Checks are usually I30572, TME 419 or other well utilized clone from the Genetic gain population. Accession plots are usually a row of 10 plants [38] although CIAT plots are a row of 12 plants [36]. Spacing is 1m between rows and plants. CETs are usually carried out in a single location. They are also unreplicated. Clone accessions are randomized. Clone accessions range from 250 to 400 mostly. However, using GS

approaches, CETs can be taken to multiple locations taking advantage of replication by families. This means that clones from a half sib or full sib family at taken to other locations and connectivity of data is achieved by using the genomic relationship matrix (GRM). This will be discussed later in detail. The selection criteria at this stage is the selection index values. Estimates of the genotypic value based on Best Linear Unbiased Prediction (BLUP) values [39] or adjusted means [40] are weighted with the economic weights set by the breeder. The selection index values are obtained as: $v^T X$ where v are the economic weights and X represents the genotypic values of clones for target traits [41-42]. The selection index provides the basis for ranking all clones in the CETs. Selected top clones then move to the next stage.

3. **Preliminary yield trials (PYTs):** The PYT design is as the CET but with reduced clone accessions (about 35 - 80) and with 2 - 3 replications. PYTs are multilocation trials with about 2 - 3 locations. Accession plots are usually 4 rows of 5 plants (20 plants; Personal communication, IKpan Smith).

4. **Advanced yield trials (AYTs):** The AYT design is as the PYT but with reduced clone accessions (about 24 - 40) and with 3 replications. AYTs are multilocation trials with about 3 - 4 locations. Accession plots are usually 6-7 rows of 6 plants (36 - 42 plants) (Personal communication, IKpan Smith).

5. **Uniform yield trials (UYTs):** The UYT design is a randomized complete block design with reduced clone accessions (about 15 - 25) and with 3 - 4 replications. UYTs are multilocation trials with about 3 - 4 locations. Accession plots are usually 6-7 rows of 6 plants (36 - 42 plants; Personal communication, IKpan Smith).

6. **On-farm trials:** Following UYTs, 2 - 5 elite clones are selected and evaluated in different regions at farmers fields. Then chosen clones are sent to the varietal release committee.

This scheme takes about 7-9 years before a variety is due for release. Elite clones are also distributed to other countries in sub-Saharan Africa using a tissue culture protocol to avoid dissemination of clones with virus contaminations [17].

1.3.4 An overview of GS:

GS is a selection technique based on genomic breeding values (GEBVs) predicted from whole genome markers (usually SNPs). GS has become feasible due to the large number of SNPs discovered by genome sequencing and new methods to efficiently genotype these SNPs [43-45]. Implementation of GS requires phenotyping and genotyping of a reference or training population [43-45]. Given this data, a prediction model that generates GEBVs from associating phenotypes to genotypes are obtained [43-45]. Subsequently, this prediction model is used to obtain GEBVs for selection candidates genotyped for SNP markers as in the reference population but not phenotyped [43,45]. Selections of new parents are now made using GEBVs of these candidates. This gain in time due to non phenotyping of candidates before selection is made, leads to shorter breeding cycles [46]. However, to maintain the accuracy of the prediction model, selected candidates need to be phenotyped and these used to update the prediction model [43-46]. GS has been shown to lead to higher gains per unit time compared to phenotypic selection in crops [46-47]. Implementation of GS has major implications for genetic evaluation systems and for genetic

improvement programmes [43-45]. These implications for cassava breeding is discussed in a later section. For complex traits that adhere to the assumptions of the infinitesimal model, the Genomic BLUP (GBLUP) model should perform well while for oligogenic traits, bayesian alphabet models that have some form of variable selection on SNPs imposed due to different priors should perform better [49-50]. However, it has been shown that accuracies of these GS prediction models are similar [51]. Higher prediction accuracies of GS models for different traits of interest are mostly due to better tracking of genetic relationship between genotypes [52].

1.3.5 GS for cassava genetic improvement:

Providing breeding value estimates from data is termed the genetic evaluation system. The type of GS prediction model used in the genetic evaluation system impacts the prediction accuracy and selection gain [51]. Genetic evaluation systems on many species have usually been carried out using the single-trait BLUP prediction model (uT) [53]. The uT GBLUP model yields GEBVs for one trait at a time ignoring information from genetically correlated traits [54]. In contrast, the multitrait BLUP prediction model (MT) accounts for genomic and residual correlations between traits when predicting GEBVs [55-56]. The genomic covariance matrix from the MT model is unstructured with diagonals as the genomic variances for traits in the analysis and the off-diagonals are genomic covariances between analyzed traits. The MT residual covariance matrix is similar with error variances on diagonals and error correlations in off-diagonals. When traits with low heritabilities are jointly analysed with high heritability traits in the MT GBLUP model, the later should benefit more from the former thereby

leading to more accurate breeding values provided that genetic correlations between these traits are significant [57]. These high heritability traits can also get little benefits from such analysis. With significant genetic correlations between traits, low residual correlations and joint analysis of high and low heritability traits; the MT model prediction accuracies are expected to be higher than those from the uT model [54,57]. However if this is not the case, the MT model does not provide any advantages over the uT model [54]. In this work (Objective 1), we tried to compare prediction accuracies from the uT and MT GBLUP models using data from the IITA collected for 16 years at three locations Ubiaja, Mokwa and Ibadan.

Another application of the MT model is for understanding the impact of GxE on the accuracies of GEBVs. When same trait at different locations are fit jointly in an MT model, we term that the multivariate multi-environment model (ME) [58-59]. The genomic covariance matrix from the ME model is unstructured and captures the genomic variance of a trait at different locations in its diagonals. The genomic covariances between a trait at analyzed locations in the off-diagonals of this matrix represent estimates of GxE. However, the error covariances from the ME model is fixed at zero reflecting an assumption that there is no residual correlation between a trait measured at multiple locations [59]. It is interesting to compare the ME model to a univariate multi-environment (uE) model. The uE model of our interest is a multikernel mixed model that fits the genomic effect as one kernel and the genomic clone-by-location effect as a second kernel. This model is equivalent to the compound symmetry (CS) model. Data for the uE model is a concatenated vector of a trait at different locations and the known covariance for the first kernel is the GRM. However the known covariance for the second kernel is a block diagonal matrix of GRMs of clones

evaluated at the different locations in the analysis. This block diagonal matrix assists in extracting the heritable genomic component of the clone-by-location interaction.

In this work, we compared the uE and ME models using 16 years multi-environment (METs) data from the IITA (Objective 1). This is the first attempt to the best of our knowledge where the uT and MT or the uE and ME models were compared in cassava. In other species, the uT and MT models have been compared [57].

A critical point where cassava breeding can take advantage of GS is at the CETs stage. For a breeding scheme based on GS, several recurrent selection cycles from the CETs to the crossing block, seedling nursery and back to the CETs can be performed. We know that prediction accuracies of GS models are also affected by genotype-by-environment (GxE) interactions [60]. However, a breeder can capture GxE impacts as early as the CETs stage using the GRM if the siblings from same family (half or full sib families) are planted across different locations. Since the GRM can track genetic relationships [52], then the GS model can better connect phenotypic information from families and also account for the effect of GxE [81]. Theoretically, this should result in more accurate breeding values for the evaluated clones. At the stage of PYT and AYT, more accurate GEBV estimates of FYLD and RTNO are obtained because replications within trials and between locations help to reduce signal-to-noise ratios and also better differentiate genomic and GxE signals. At this stage, a good clone can still be sent back to the crossing block to generate superior progenies. In the next section, we delve into mapping complex traits in African cassava via methods based on genomic segments instead of SNPs.

1.3.6 Mapping complex traits in cassava using the Regional Heritability Mapping (RHM) or the Regional co-heritability Mapping (RcHM) procedures:

Genetic mapping of complex traits has been an interesting arena in the field of quantitative genetics. Plant populations harbour a diversity of phenotypic variation for morphology, physiology, behaviour, performance and disease susceptibility. This observed variation is due to an underlying genetic complexity from interaction of many loci, with allelic effects that are sensitive to the environmental cues an individual is exposed to [61-62]. The principles of mapping quantitative trait loci (QTL) that affect variation in complex traits are known [63]. This is based on assumed linkage of segregating polymorphic genetic markers with underlying QTLs affecting traits of interest [63-64]. Mapping QTLs has two components: detection and localization [64]. The power to detect QTL affecting a complex trait depends on their effects and allele frequencies [64]. The effect is the average difference in the phenotype between marker allele genotypes scaled by the phenotypic standard deviation of the trait within marker genotype classes [64]. QTL mapping takes advantage of recent recombination events although with large blocks of local linkage disequilibrium (LD) in an F2 or backcross population [62]. However with the advent of genotyping technologies that can type many SNPs across the genome, a divergent population with different distributions of global LD (due to ancient recombinations and coalescence) across the genome can be used to associate QTLs to a trait with better precision on the localization of the underlying QTL [64]. This procedure is known as genome-wide association analysis (GWAS) [65]. GWAS has been successful in mapping some important cassava traits [66]. GWAS is based on

single-SNPs associations which have been shown to be powerful at capturing common variants associated with traits of interest [65,67]. However, some studies have utilized multi-SNP association approaches based on haplotypes for understanding the genetic basis of complex traits [68-69]. As an inheritance unit and a form of genetic variation, a haplotype like SNPs may affect phenotypes either through influencing promoters and protein structure [70-71] or by tagging nearby untyped or rare causal variants [72-73]. This makes haplotype association of great interest for unraveling the genetic basis of complex traits [67]. Multi-SNP association approaches take into account two types of heterogeneity that are blind spots for the single-SNP GWAS analysis [67]. These include allelic heterogeneity - a situation where different mutations within a gene cause a similar phenotype or locus heterogeneity - where mutations at different genes cause a similar phenotype [67]. These make multi-SNP association approaches more powerful than GWAS [67]. Also multi-SNP association approaches account for possible interactions among SNP markers [67]. Another multi-SNP association approach which is markedly different from the haplotype association analysis is the RHM [74-75].

The differences between haplotype association analysis and RHM are detailed below:

1. Haplotype analysis relies on complex powerful models called Hidden Markov Models to infer ancestral haplotypes from SNP data [67,76-77]. These algorithms have been shown to be accurate with the PHASE software being a standard in this arena [76-77]. However after reconstructing haplotypes from data, haplotypes need to be grouped to avoid testing a large number of haplotypes [67] which will decrease the degrees of free-

dom for the test statistic. However for RHM, a sliding window approach based on a given number of SNPs is used to construct genomic segments along the genome [74-75]. The number of SNPs within the window may be arbitrary but information on the local LD along the chromosome can be used to construct genomic segments or regions. These segments may represent pseudo haplotypes.

2. Haplotypes are multi-allelic but still need to be grouped and coded into different genotype classes [67]. However for RHM segments, a GRM is calculated at each segment and these help track the genetic relationship between genotypes at target segments. This represents a key difference between haplotypes and RHM segments and is expected to influence the results of both analysis.

It is expected that both haplotype and RHM methods can capture rare variants and QTL interactions that affect complex traits [74-75, 78]. Hence results from both analyses are not expected to be markedly different.

We also developed an RcHM analysis akin to the RHM. The RcHM analysis differs from the RHM in the dimension of traits. The RcHM attempts to map the co-inheritance of two traits at each segment in the genome. An ideal RcHM analysis slides through the genome in windows and fits a bivariate model with two kernels representing the target genomic segment and the rest of the genome respectively. The later accounts for background effects in the genome or corrects for population structure. However due to convergence problems, an alternative RcHM was developed in objective 4 of this work which fits a bivariate SNP-BLUP model [79] first and then sums of the effects of SNPs in a segment to obtain genomic segment values (GSVs) for both traits. Subsequently, a pair-

wise correlation of GSVs for both traits at each segment yields genomic segment correlations which represent the co-inheritance of both traits at each segment. A genome-wide map of genomic segment correlations would reveal the co-inheritance profile of two traits across the entire genome. This information will allow the breeder to understand the association of two traits at all segments in the genome thus providing a vital information that may affect how both traits are genetically improved.

1.4 Dissertation organization:

The second chapter deals with the comparison of multivariate and univariate GS models via prediction accuracies in African cassava. The third chapter addresses the genetic basis of cassava DM in two subpopulations of cassava using the RHM procedure. The fourth chapter provides insights into the co-inheritance of DM, tuber yellowness and FYLD in different cassava subpopulations using the RcHM procedure. The fifth and final chapter provides a conclusion for the dissertation.

1.5 References:

[1] El-Sharkawy, M.A., 2003. Cassava biology and physiology. Plant molecular biology, 53(5), pp.621-641.

[2] Keating, B.A., Wilson, G.L. and Evenson, J.P., 1988. Effects of length, thickness, orientation, and planting density of cassava (Manihot esculenta Crantz) planting material on subsequent establishment, growth and yield. E. Afr. Agric. For. J, 53, pp.145-149.

[3] Cock, J.H., 1982. Cassava: a basic energy source in the tropics. Science, 218(4574), pp.755-762.

[4] Shore, K., 2002. Decades of cassava research bear fruit. IDRC reports, Apr. 26, 2002.

[5] Kawano, K., Fukuda, W.M.G. and Cenpukdee, U., 1987. Genetic and environmental effects on dry matter content of cassava root. Crop Science, 27(1), pp.69-74.

[6] Holleman, L.W.J., and A. Aten. 1956. Elaboracion de la yuca y sus productos en las industrias rurales. Cuaderno de Fomento Agropecuario. Organizacion de las Naciones Unidas para la Agricultura y la Alimentacion. Bol. 54.

[7] Barrios, E.A., and R. Bressani. 1967. Composicion quimica de la raiz y de la hoja de algunas variedades de yuca Manihot. Turrialba 17:314-320.

[8] Lim, H.K. 1968. Composition data of feeds and concentrates. Malay.Agric.J.46:63-79.

[9] Tonukari, N.J., 2004. Cassava and the future of starch. Electronic journal of biotechnology, 7(1), pp.5-8.

[10] Srinivas, T., 2007. Industrial demand for cassava starch in India. Starch-Strke, 59(10), pp.477-481.

[11] Ceballos, H., Iglesias, C.A., Prez, J.C. and Dixon, A.G., 2004. Cassava breeding: opportunities and challenges. Plant molecular biology, 56(4), pp.503-516.

[12] Ceballos, H., Fregene, M., Prez, J.C., Morante, N. and Calle, F., 2007. Cassava genetic improvement. Breeding major food staples, pp.365-391.

[13] Sayre, R., Beeching, J.R., Cahoon, E.B., Egesi, C., Fauquet, C., Fellman, J., Fregene, M., Gruissem, W., Mallowa, S., Manary, M. and Maziya-Dixon, B., 2011. The BioCassava plus program: biofortification of cassava for sub-Saharan Africa. Annual review of plant biology, 62, pp.251-272.

[14] Saltzman, A., Birol, E., Bouis, H.E., Boy, E., De Moura, F.F., Islam, Y. and Pfeiffer, W.H., 2013. Biofortification: progress toward a more nourishing future. Global Food Security, 2(1), pp.9-17.

[15] Beck, B.D.A., 1982. Historical perspectives of cassava breeding in Africa. In Root crops in Eastern Africa: proceedings of a workshop held in Kigali, Rwanda, 23-27 Nov. 1980. IDRC, Ottawa, ON, CA.

[16] Jennings, D.L. and Hershey, C.H., 1985. Cassava breeding: a decade of progress from international programmes. Progress in plant breeding, 1, pp.89-115.

[17] Manyong, V.M., 2000. Impact: The Contribution of IITA-improved Cassava to Food Security in Sub-Saharan Africa. IITA.

[18] Prochnik, S., Marri, P.R., Desany, B., Rabinowicz, P.D., Kodira, C., Mohiuddin, M., Rodriguez, F., Fauquet, C., Tohme, J., Harkins, T. and Rokhsar, D.S., 2012. The cassava genome: current progress, future directions. Tropical plant biology, 5(1), pp.88-94.

[19] Heffner, E.L., Lorenz, A.J., Jannink, J.L. and Sorrells, M.E., 2010. Plant breeding with genomic selection: gain per unit time and cost. Crop science, 50(5), pp.1681-1690.

[20] Ceballos, H., Kawuki, R.S., Gracen, V.E., Yencho, G.C. and Hershey, C.H., 2015. Conventional breeding, marker-assisted selection, genomic selection and inbreeding in clonally propagated crops: a case study for cassava. Theoretical and Applied Genetics, 128(9), pp.1647-1667.

[21] Nassar, N. and Ortiz, R., 2010. Breeding cassava to feed the poor. Scientific American, 302(5), pp.78-84.

[22] Awoleye F, Duren M, Dolezel J et al (1994) Nuclear DNA content and in vitro induced somatic polyploidization cassava (Manihot esculenta Crantz) breeding. Euphytica 76:195202.

[23] Wang, W., Feng, B., Xiao, J., Xia, Z., Zhou, X., Li, P., Zhang, W., Wang, Y., Mller, B.L., Zhang, P. and Luo, M.C., 2014. Cassava genome from a wild ancestor to cultivated varieties. Nature communications, 5.

[24] Hahn, S.K., Terry, E.R. and Leuschner, K., 1980. Breeding cassava for resistance to cassava mosaic disease. Euphytica, 29(3), pp.673-683.

[25] HAHN, S. K., 1978. Breeding cassava for resistance to bacterial blight. PANS 24(4): 480485.

[26] HAHN, S. K., A. K. HOWLAND and E. R. TERRY, 1980. Correlated resistance of cassava to mosaic and bacterial blight diseases. Euphytica 29 : 305-3 11.

[27] HAHN, S. K., A. K. HOWLAND and E. R. TERRY, 1973. Cassava breeding at IITA. Proc. Int. Trop. Root Crop Symp., Ibadan, Nigeria. pp. 410.

[28] Kawuki, R.S., Kaweesi, T., Esuma, W., Pariyo, A., Kayondo, I.S., Ozimati, A., Kyaligonza, V., Abaca, A., Orone, J., Tumuhimbise, R. and Nuwamanya, E., 2016. Eleven years of breeding efforts to combat cassava brown streak disease. Breeding science, 66(4), pp.560-571.

[29] Nweke, F.I., 1993. Cassava varietal needs of farmers and the potential for production growth in Africa. COSCA Working Paper, 10.

[30] Nweke, F.I., Spencer, D.S. and Lynam, J.K., 2002. The cassava transformation: Africa's best-kept secret. Michigan State University Press.

[31] Enete, A., Nweke, F. and Tollens, E., 2002. Contributions of men and women to food crop production labour in Africa: information from COSCA. Outlook on Agriculture, 31(4), pp.259-265.

[32] Gutierrez, A.P., Yaninek, J.S., Wermelinger, B., Herren, H.R. and Ellis, C.K., 1988. Analysis of biological control of cassava pests in Africa. III. Cassava green mite Mononychellus tanajoa. Journal of Applied Ecology, pp.941-950.

[33] Polson, R.A. and Spencer, D.S., 1991. The technology adoption process in

subsistence agriculture: The case of cassava in Southwestern Nigeria. Agricultural systems, 36(1), pp.65-78.

[34] Bouis, H.E., Hotz, C., McClafferty, B., Meenakshi, J.V. and Pfeiffer, W.H., 2011. Biofortification: a new tool to reduce micronutrient malnutrition. Food and nutrition bulletin, 32(1 suppl1), pp.S31-S40.

[35] Kulakow, P. and Parkes, E., 2015. Vitamin A Cassava. trials, 100(4), pp.1734-1746.

[36] Kawano, K., 2003. Thirty years of cassava breeding for productivitybiological and social factors for success. Crop Science, 43(4), pp.1325-1335.

[37] Kawuki, R.S., Pariyo, A., Amuge, T., Nuwamanya, E., Ssemakula, G., Tumwesigye, S., Bua, A., Baguma, Y., Omongo, C., Alicai, T. and Orone, J., 2011. A breeding scheme for local adoption of cassava (Manihot esculenta Crantz). Journal of Plant Breeding and Crop Science, 3(7), pp.120-130.

[38] International Institute of Tropical Agriculture, 1990. Cassava in Tropical Africa: A Reference Manual. IITA.

[39] Henderson, C.R., 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics, pp.423-447.

[40] Spilke, J., Piepho, H.P. and Hu, X., 2005. Analysis of unbalanced data by mixed linear models using the MIXED procedure of the SAS system. Journal of Agronomy and Crop Science, 191(1), pp.47-54.

[41] Lin, C.Y., 1978. Index selection for genetic improvement of quantitative characters. TAG Theoretical and Applied Genetics, 52(2), pp.49-56.

[42] Smith, H.F., 1936. A discriminant function for plant selection. Annals of Human Genetics, 7(3), pp.240-250.

[43] Goddard, M.E. and Hayes, B.J., 2007. Genomic selection. Journal of Animal breeding and Genetics, 124(6), pp.323-330.

[44] Hayes, B.J., Bowman, P.J., Chamberlain, A.J. and Goddard, M.E., 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. Journal of dairy science, 92(2), pp.433-443.

[45] Meuwissen, T., 2007. Genomic selection: marker assisted selection on a genome wide scale. Journal of animal Breeding and genetics, 124(6), pp.321-322.

[46] Heffner, E.L., Sorrells, M.E. and Jannink, J.L., 2009. Genomic selection for crop improvement. Crop Science, 49(1), pp.1-12.

[47] Lorenzana, R.E. and Bernardo, R., 2009. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theoretical and applied genetics, 120(1), pp.151-161.

[48] Wimmer, V., Lehermeier, C., Albrecht, T., Auinger, H.J., Wang, Y. and Schn, C.C., 2013. Genome-wide prediction of traits with different genetic architecture through efficient variable selection. Genetics, 195(2), pp.573-587.

[49] Daetwyler, H.D., Pong-Wong, R., Villanueva, B. and Woolliams, J.A., 2010. The impact of genetic architecture on genome-wide evaluation methods. Genetics, 185(3), pp.1021-1031.

[50] Gianola, D., de los Campos, G., Hill, W.G., Manfredi, E. and Fernando,

R., 2009. Additive genetic variability and the Bayesian alphabet. Genetics, 183(1), pp.347-363.

[51] Heslot, N., Yang, H.P., Sorrells, M.E. and Jannink, J.L., 2012. Genomic selection in plant breeding: a comparison of models. Crop Science, 52(1), pp.146-160.

[52] Habier, D., Fernando, R.L. and Dekkers, J.C.M., 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics, 177(4), pp.2389-2397.

[53] Goddard, M., 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica, 136(2), pp.245-257.

[54] Guo, G., Zhao, F., Wang, Y., Zhang, Y., Du, L. and Su, G., 2014. Comparison of single-trait and multiple-trait genomic prediction models. BMC genetics, 15(1), p.30.

[55] Van der Werf, J.H.J., Van Arendonk, J.A.M. and De Vries, A.G., 1992. Improving selection of pigs using correlated characters.

[56] Ducrocq, V., 1994. Multiple trait prediction: principles and problems. 5th World Congr Genet Appl Livest Prod, Guelph, pp.7-12.

[57] Jia, Y. and Jannink, J.L., 2012. Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics, 192(4), pp.1513-1522.

[58] Falconer, D.S., 1952. The problem of environment and selection. The American Naturalist, 86(830), pp.293-298.

[59] Burgueo, J., de los Campos, G., Weigel, K. and Crossa, J., 2012. Genomic

prediction of breeding values when modeling genotype environment interaction using pedigree and dense molecular markers. Crop Science, 52(2), pp.707-719.

[60] Ly, D., Hamblin, M., Rabbi, I., Melaku, G., Bakare, M., Gauch Jr, H.G., Okechukwu, R., Dixon, A.G., Kulakow, P. and Jannink, J.L., 2013. Relatedness and genotype environment interaction affect prediction accuracies in genomic selection: a study in cassava. Crop Science, 53(4), p.1312.

[61] TFC, F.D.M., 1996. Introduction to quantitative genetics. New York, NY: Pearson/Prenctice Hall, 463, p.464.

[62] Lynch, M. and Walsh, B., 1998. Genetics and analysis of quantitative traits (Vol. 1, pp. 535-557). Sunderland, MA: Sinauer.

[63] Sax, K., 1923. The association of size differences with seed-coat pattern and pigmentation in Phaseolus vulgaris. Genetics, 8(6), p.552.

[64] Mackay, T.F., Stone, E.A. and Ayroles, J.F., 2009. The genetics of quantitative traits: challenges and prospects. Nature reviews. Genetics, 10(8), p.565.

[65] Korte, A. and Farlow, A., 2013. The advantages and limitations of trait analysis with GWAS: a review. Plant methods, 9(1), p.29.

[66] Wolfe, M.D., Rabbi, I.Y., Egesi, C., Hamblin, M., Kawuki, R., Kulakow, P., Lozano, R., Carpio, D.P.D., Ramu, P. and Jannink, J.L., 2016. Genome-wide association and prediction reveals genetic architecture of cassava mosaic disease resistance and prospects for rapid genetic improvement. The plant genome, 9(2).

[67] Xu, H. and Guan, Y., 2014. Detecting local haplotype sharing and haplotype association. Genetics, pp.genetics-114.

[68] Trgout, D.A., Knig, I.R., Erdmann, J., Munteanu, A., Braund, P.S., Hall, A.S., Grohennig, A., Linsel-Nitschke, P., Perret, C., DeSuremain, M. and Meitinger, T., 2009. Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. Nature genetics, 41(3), pp.283-285.

[69] GChang, M.C., Chang, Y.T., Tien, Y.W., Liang, P.C., Jan, I.S., Wei, S.C. and Wong, J.M., 2007. T-cell regulatory gene CTLA-4 polymorphism/haplotype association with autoimmune pancreatitis. Clinical Chemistry, 53(9), pp.1700-1705.

[70] Drysdale, C.M., McGraw, D.W., Stack, C.B., Stephens, J.C., Judson, R.S., Nandabalan, K., Arnold, K., Ruano, G. and Liggett, S.B., 2000. Complex promoter and coding region 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. Proceedings of the National Academy of Sciences, 97(19), pp.10483-10488.

[71] Joosten, P.H., Toepoel, M., Mariman, E.C. and Van Zoelen, E.J., 2001. Promoter haplotype combinations of the platelet-derived growth factor [alpha]-receptor gene predispose to human neural tube defects. Nature genetics, 27(2), p.215.

[72] Clark, A.G., 2004. The role of haplotypes in candidate gene studies. Genetic epidemiology, 27(4), pp.321-333.

[73] Servin, B. and Stephens, M., 2007. Imputation-based analysis of associa-

tion studies: candidate regions and quantitative traits. PLoS genetics, 3(7), p.e114.

[74] Riggio, V., Matika, O., Pong-Wong, R., Stear, M.J. and Bishop, S.C., 2013. Genome-wide association and regional heritability mapping to identify loci underlying variation in nematode resistance and body weight in Scottish Blackface lambs. Heredity, 110(5), p.420.

[75] Nagamine, Y., Pong-Wong, R., Navarro, P., Vitart, V., Hayward, C., Rudan, I., Campbell, H., Wilson, J., Wild, S., Hicks, A.A. and Pramstaller, P.P., 2012. Localising loci underlying complex trait variation using regional genomic relationship mapping. PloS one, 7(10), p.e46501.

[76] Stephens, M. and Donnelly, P., 2003. A comparison of bayesian methods for haplotype reconstruction from population genotype data. The American Journal of Human Genetics, 73(5), pp.1162-1169.

[77] Stephens, M., Smith, N.J. and Donnelly, P., 2001. A new statistical method for haplotype reconstruction from population data. The American Journal of Human Genetics, 68(4), pp.978-989.

[78] Cantor, R.M. and Wilcox, M., 2011. Detecting rare variant associations: methods for testing haplotypes and multiallelic genotypes. Genetic epidemiology, 35(S1).

[79] Strandn, I. and Garrick, D.J., 2009. Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. Journal of dairy science, 92(6), pp.2971-2975.

[80] 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. and Nauluvula, P., 2016. Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. Nature biotechnology, 34(5), pp.562-570.

[81] Endelman, J.B., Atlin, G.N., Beyene, Y., Semagn, K., Zhang, X., Sorrells, M.E. and Jannink, J.L., 2014. Optimal design of preliminary yield trials with genome-wide markers. Crop Science, 54(1), pp.48-59.

CHAPTER 2

ACCURACIES OF UNIVARIATE AND MULTIVARIATE GENOMIC PREDICTION MODELS IN AFRICAN CASSAVA.

2.1 Abstract

Genomic selection (GS) promises to accelerate genetic gain in plant breeding programs especially for crops like cassava that have long breeding cycles. To practically implement GS in cassava breeding, it is useful to evaluate different GS models and to develop suitable models for an optimized breeding pipeline. In this paper we: (1) compared prediction accuracies from a single-trait (uT) and a multi-trait (MT) mixed model for a single-environment genetic evaluation (Scenario 1), (2) compared accuracies from a compound symmetric multi-environment model (uE) parameterized as a univariate multi-kernel model and a multivariate (ME) multi-environment mixed model that each accounts for genotype-by-environment interaction for multi-environment genetic evaluation (Scenario 2). We used sixteen years of public cassava breeding data for six target cassava traits for these analyses. A 5-fold cross validation scheme with 10-repeat cycles was used to assess model prediction accuracies. In Scenario 1, the MT models had higher prediction accuracies than the uT models for most traits and locations analyzed, amounting to 40% better prediction accuracy on average. For Scenario 2, we observed that the ME model had on average (across all locations and traits) 12% better predictive ability than the uE model. We recommend the use of multivariate mixed models (MT and ME) for cassava genetic evaluation. These models may be useful for other plant species.

2.2 Keywords

Genomic selection, plant genetic evaluation, cassava breeding, single trait models, univariate multi-environment models, multi-trait (MT) models, multivariate multi-environment (ME) models, GxE interactions, prediction accuracies, Genomic estimated breeding values (GEBVs).

2.3 Background

Cassava (Manihot esculenta Crantz)[1] is a staple food for over 700 million people in Africa, South America and Asia [2]. Cassava also has immense industrial potential. White cassava starch is easy to extract and contains low levels of fat (about 1.5%), protein (about 0.6%) and phosphorus (about 4%), which are desirable attributes for the food industry [3,4]. Given the issues of climate change and rapid population growth in countries that rely heavily on cassava, rapid genetic improvement of cassava is critically needed. To enable rapid genetic improvement of cassava, genetic evaluation protocols based on Best Linear Unbiased Prediction (BLUP) analysis [5,6] and selection on a merit index [7,8] have been recommended [9] to maximize gain from selection.

Genomic selection (GS) [10] offers crops like cassava tremendous opportunity for accelerated genetic gains [11] by making use of whole genome SNP markers scored with methods like the genotyping-by-sequencing (GBS) [12]. These whole genome SNP markers could be dense enough to be in linkage disequilibrium with most quantitative trait loci (QTL) affecting traits of interest. Using GS, selection is imposed at these QTL without actually identifying the

QTL or the functional polymorphisms [10]. Also these markers will help to better track relatedness [24]. This yields an improvement in selection accuracies especially where pedigree records are not fully available [25].

GS models for plant genetic evaluation: Genetic evaluation [9] starts with accurately estimating the genetic value of an individual for a wide range of traits using its own performance records, progeny performance records, records from relatives, or a combination of the three [13]. This estimation has usually been carried out using single trait (uT) BLUP methodology [14] for obtaining estimated breeding values (EBVs) for one trait at a time. In plant and animal breeding, breeders usually select on the basis of multiple traits that are often genetically correlated. The uT model for traits measured in a single environment assumes zero genetic and residual covariances between these traits such that information from other traits are not utilized when obtaining EBVs of the evaluated individuals for the traits in the analysis. However, the optimal estimation procedure to combine information from multiple trait records and obtain EBVs is the multi-trait BLUP methodology (MT) [15,16]. The MT model does not assume zero genetic and residual covariances but rather estimates them and also uses this information when obtaining individual EBVs for the traits in the analysis. The MT model has several advantages over the uT model including:

- Higher prediction accuracies for individual traits in the model because of more information (direct or indirect) and better connectedness of the data [17], especially when traits with varying heritabilities are analyzed jointly. This is true if the genetic correlations in the model are significant or substantial.
- Simplified index selection because optimal weight factors for the total

merit index are the economic weights [17].

- Procedures for obtaining genetic and residual covariances and incorporating these into EBV estimates for across-location, -country or -region evaluations [18, 19].
- Better selection accuracies when all target traits under selection are included in the model [20] in addition to the use of all individuals (selected or not) in the relationship matrix.

While MT models have clear advantages over uT models they require the estimation of additional parameters (i.e., the genetic and error covariances), which will affect accuracies of EBVs. The number of additional parameters increases as the number of traits increases. For large models, many additional parameters can lead to convergence problems in the analysis. Lastly, an appreciable amount of data is required to get good estimates of these additional parameters.

In most plant breeding programs, genotypes are evaluated in multi-environment trials (METs) usually at advanced stages of breeding. The goal is to sample the influence on selection candidates of the range of environments for which varieties will be targeted. Addressing the problem of the analysis of METs brings into focus another potential use for MT models [30]. Here, phenotypes of the same trait, but measured at different locations are parameterized as different traits in the MT model [31], producing what we call a multi-environment BLUP (ME) model. Like the MT model, the ME model estimates genetic covariances between a single trait measured at multiple environments which may lead to more accurate estimates of individual EBVs for the trait at all the environments where data has been recorded. For ME models used for modelling MET data, residual covariances are set to zero reflecting the assump-

tion that no mechanism generates error covariances between a trait measured in different environments [18]. In contrast, the typical univariate BLUP model for modelling METs data, termed the univariate multi-environment model (uE), fits a multi-kernel mixed model with the genotypic effect as one kernel and the genotype-by-environment (GxE) effect as the second kernel and maybe environment as third kernel [26]. This model yields a GxE variance for a MET and individuals can be ranked on their performance at different locations. Different variants of the ME model have been used for modeling environment covariance structures in plant [32-35] and in animal breeding [36,37]. Genetic covariances from the ME model offer a convenient tool for assessing the impact of GxE on a trait. The genetic covariances relate directly to the extent of GxE at all locations in the analysis. A low genetic correlation of the EBVs between a trait at different locations from the ME model indicates high GxE impact on that trait [9, 38-41].

Selecting the GS model to be employed in a practical cassava breeding program requires comparing models that will be useful in the different stages of cassava breeding with METs data. Finally, fitting multivariate BLUP models is not trivial. Even with software that can in principle fit these models, model convergence is not guaranteed and may require several attempts [21-23] such that univariate models may be more practical if benefits of the multivariate models are not substantial.

The objectives of this paper are to:

- Compare multi-trait (MT) and single trait (uT) mixed models for single environment data using cross-validated prediction accuracies.
- Compare the multivariate multi-environment (ME) model to a single-trait multi-environment (uE) model using cross-validated prediction accuracies and assessing GxE impact on analyzed traits via genetic covariances from the ME model fit.

2.4 Materials and Methods

2.4.1 Cassava phenotype data:

We used historical phenotype data from different trials conducted by the cassava breeding program at the International Institute of Tropical Agriculture (IITA), Ibadan, Nigeria in our analysis. The Genetic Gain population represents a collection of clones selected from the 1970s to 2007 by the cassava breeding program at the IITA [48,49]. Some of these clones are West African landraces and some are of East African origin. Clones in the Genetic Gain population have gone through advanced stages of the cassava breeding process up to on-farm variety testing trials. The data used in our analysis comprises data collected on clonal evaluation trials (CETs) which are augmented design trials with typically 2 known checks and unreplicated plots with 5 plants. These data were collected from three target locations in Nigeria: Ibadan (7.40 N, 3.90 E), Mokwa (9.3 N, 5.0 E), and Ubiaja (6.66 N, 6.38 E). These locations represent regions which encompass about 35% of the cassava production base in Nigeria. Data sets were

collected from 2000 to 2015 and included trials with most of the 739 clones of the Genetic Gain population. Six target agronomic traits were used in the analysis including seedling vigor (VIGOR), Number of storage roots per plot at harvest (RTNO), Fresh weight of harvested roots expressed in tons per hectare (T/ha) (FYLD), percent dry matter (DM) of storage roots, which measures root dry weight as the % of the root fresh weight, plot mean cassava mosaic disease severity (MCMDS), rated on a scale from 1 (no symptoms) to 5 (extremely severe), and plot mean cassava green mite (MCGM) severity, rated on a scale from 1 (no symptoms) to 5 (extremely severe). Cassava mosaic disease is caused by a Begomovirus that belongs to the *Geminiviridae* family, and is carried and transmitted by the whitefly *Bemisia tabaci*. The cassava green mite is *Mononychellus tanajoa* [50]. These traits are target traits used in the selection index for selection decisions in the IITA cassava breeding program. Phenotype data metrics are shown in Table 2.1. All trait records were plot averages for both clonal accessions and checks. All checks were included in the analysis.

Table 2.1: **Cassava phenotype means and standard deviations (in braces) at 3 locations: Ubiaja, Mokwa, and Ibadan.**

	Ubiaja	Mokwa	Ibadan
No. of records	7806	5345	5579
No. of Clones	739	573	691
VIGOR	6.51 (1.12)	6.52 (0.93)	6.11 (1.23)
RTNO	31.71 (17.30)	37.05 (21.76)	37.87 (23.91)
FYLD	12.61 (7.70)	16.51 (9.54)	15.84 (10.72)
DM	31.95 (6.42)	29.01 (6.38)	30.8 (6.79)
MCMDS	1.59 (0.93)	1.21 (0.57)	2.14 (1.01)
MCGM	3.56 (0.97)	2.99 (0.67)	3.00 (0.85)

Traits are: VIGOR (seedling vigor), RTNO (Number of storage roots per plot), FYLD (fresh weight of harvested roots in tons per hectare), DM (percentage dry matter in roots), MCMDS (plot mean cassava mosaic disease severity) and MCGM (plot mean cassava green mites severity).

2.4.2 Cassava genotype data:

DNA was extracted using DNeasy Plant Mini Kits (Qiagen) from 739 clones from the 2013 Genetic Gain trial at IITA and was quantified using PicoGreen. Genotyping-by-sequencing (GBS) was used for genotyping [12] these clones. Six 95-plex and one 75-plex ApeKI libraries were constructed and sequenced on Illumina HiSeq, one lane per library. Single nucleotide polymorphisms (SNPs) were called from the sequence data using the TASSEL pipeline version 4.0 [51], using an alignment to the *Manihot esculenta* version 6 reference genome [52]. Average sequencing depth for polymorphic loci was 5x. Individuals with greater than 80% and markers with more than 60% missing calls were removed. The

marker data was converted to dosage format (0, 1, 2) and missing genotypic data were imputed using a LASSO regression method (Ariel Chan, personal communication, 2014) implemented using the R glmnet package [53]. The final data set consisted of 183,201 SNPs scored in 739 clones.

2.4.3 Statistical analysis:

We structured the cassava phenotype data described above into two types of data common in most plant breeding programs. The first set was achieved by pooling data from multiple years at specific locations (multi-year trials data). We termed this scenario the single-environment genetic evaluation (Scenario 1). The resulting predictive ability from this data were assessed for the three locations. The second scenario was achieved by using data from multiple locations and years (METs) but in this case extracting location specific information by modeling GxE interaction. We termed this scenario the multi-environment genetic evaluation (Scenario 2). The goal of the latter scenario is to assess the value of evaluating the impact of GxE and check if this yields better predictive value of the breeding value of a clone.

2.4.4 Pseudo-true genetic values for model accuracy computations:

For validating the models in this study, we define first a univariate single trait mixed model for each trait at each location separately (to preserve the variation embedded in each location) using an identity covariance matrix among clone effects, which assumes no relationship among all clones. The univariate mixed model was as follows:

$$\begin{gathered} \mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \\ \mathbf{u} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I}); \quad \mathbf{e} \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I}) \end{gathered} \tag{2.1}$$

Where **y** is a vector of observations, **b** is a vector of fixed effects with design matrix **X** (relating observations to fixed effects in this case including grand mean and a nested effect of Trial-within-Year and the ratio of plants harvested to number planted); **u** is a vector of clonal genetic effects with design matrix **Z** (relating observations to clones). This model was fit using the lmer function in the R lme4 package [55] and resulting BLUP values $\hat{\mathbf{u}}$, which we refer to as Estimated Genotypic Values (EGVs), were used as pseudo-true genetic effects for prediction accuracy computations. This follows practice in the plant breeding literature [50, 70-71].

2.4.5 GS models for Scenario 1:

We define two mixed models fitted here as follows:

2.4.6 The single trait mixed model (uT):

$$\begin{gathered} \mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e} \\ \mathbf{u} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{K}); \quad \mathbf{e} \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I}) \end{gathered} \tag{2.2}$$

Where $\mathbf{y}$ is the response vector of a trait for a location, β is the vector of fixed effects with design matrix $\mathbf{X}$ (relating observations to fixed effects namely the grand mean, nesting of Trial-within-Year and ratio of plants harvested to number planted); $\mathbf{u}$ is the vector of random additive genomic effects with design matrix $\mathbf{Z}$ (relating trait values to clones) and $\mathbf{K}$ is the additive genomic relationship matrix generated from SNP markers as in method 1 of VanRaden, 2008 [63] implemented in preGSf90 [62].

2.4.7 The multitrait mixed model (MT):

$$\begin{gathered} \mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e} \\ \mathbf{y} = (\mathbf{y}_1^{'}, \mathbf{y}_2^{'}, \mathbf{y}_3^{'}, \mathbf{y}_4^{'}, ..., \mathbf{y}_d^{'}); \quad u = (\mathbf{u}_1^{'}, \mathbf{u}_2^{'}, ..., \mathbf{u}_d^{'}); \\ \mathbf{e} = (\mathbf{e}_1^{'}, \mathbf{e}_2^{'}, \mathbf{e}_3^{'}, \mathbf{e}_4^{'}, ..., \mathbf{e}_d^{'}); \end{gathered} \tag{2.3}$$

Response $\mathbf{y}$ is a vector of d traits (six core traits described above) in locations a, b and c (corresponding to Ubiaja, Mokwa and Ibadan) recorded for n clones, $\mathbf{X}$ and $\mathbf{Z}$ were design matrices as $\mathbf{Xa}$, $\mathbf{Xb}$ or $\mathbf{Xc}$ and $\mathbf{Za}$, $\mathbf{Zb}$ or $\mathbf{Zc}$ respectively for for fixed effects β (with components as in model 2 above for every location and trait) and random genetic effects $\mathbf{u}$ for the locations a, b and c allowing for

missing clones and observations. Following a multivariate normal distribution (Nm), the marginal density of y is given as:

$$\begin{aligned} (\mathbf{y}\,|\,\beta, \mathbf{R}, \mathbf{G}) &\sim N_m(\mathbf{X}\beta, \mathbf{V}) \\ var(\mathbf{y}) = \mathbf{V} = \mathbf{Z}(\mathbf{G} \otimes \mathbf{K})\mathbf{Z}^T + \mathbf{R} \otimes \mathbf{I}); \quad \hat{\mathbf{u}} &= (\mathbf{G} \otimes \mathbf{K})\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta) \end{aligned} \tag{2.4}$$

The matrices **G** and **R** are $d \times d$ symmetric unstructured genomic and error covariance matrices respectively, **K** remains the additive genomic relationship matrix for n clones generated from SNP markers as above, **I** is an identity matrix and **u** are the genomic estimated breeding values (GEBVs) of the clones for the traits in the analysis.

Models (2.2) and (2.3) were fitted separately for the locations Ubiaja, Mokwa and Ibadan respectively, allowing the error (co)variances associated with these locations to be distinct. Note also that genotype-by-location effects are confounded with main genotype effects in these models such that variance components may change between locations. The effects of years and trials were fixed because our emphasis was on location effects as these locations represented different production regions and we sought to capture consistent effects of these locations. In contrast, year effects are variable and by definition not consistent. Also following practice in cassava breeding [70-71], multiple observations of a clone were not considered as repeated measures. Although these subjects were genetic clones, data was collected from distinct individuals making them independent. Hence these measurements were treated as samples of clones and should lead to better precision in prediction of breeding values.

2.4.8 GS models for Scenario 2:

We also defined two mixed models here with the aim of modeling genotype-by-environment interaction effects as follows:

2.4.9 The compound symmetric multi-environment model (uE):

We describe the uE model the way it is fit first, then show its compound symmetry structure. The model is as follows:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\beta + \mathbf{Z_1}\mathbf{u} + \mathbf{Z_2}\mathbf{w} + \mathbf{e} \\ \mathbf{y} &= (\mathbf{y}_a^{'}, \mathbf{y}_b^{'}, \mathbf{y}_c^{'})^{'} \quad u = (\mathbf{u}_a^{'}, \mathbf{u}_b^{'}, \mathbf{u}_c^{'})^{'}; \quad \mathbf{e} = (\mathbf{e}_a^{'}, \mathbf{e}_b^{'}, \mathbf{e}_c^{'})^{'} \\ \mathbf{u} &\sim N(\mathbf{0}, \sigma_u^2\mathbf{K}); \quad \mathbf{w} \sim N(\mathbf{0}, \sigma_w^2\mathbf{I_3} \otimes \mathbf{K}); \quad \mathbf{e} \sim N(\mathbf{0}, \sigma_e^2\mathbf{I}) \\ var(y) &= \mathbf{V} = \sigma^2 u(\mathbf{Z_1}\mathbf{K}\mathbf{Z_1}^{'}) + \sigma^2 w(\mathbf{Z_2}\mathbf{K}\mathbf{Z_2}^{'}) + \sigma^e\mathbf{I} \\ \mathbf{Z_2} &= diag(\mathbf{Z_a}, \mathbf{Z_b}, \mathbf{Z_c}) \end{aligned} \tag{2.5}$$

Where $\mathbf{y}$ is a vector of a trait at locations a, b and c (corresponding to Ubiaja, Mokwa and Ibadan), β is the vector of fixed effects with design matrix $\mathbf{X}$ (relating observations to fixed effects as in model 2); $\mathbf{I}$ is an identity matrix and $\mathbf{I_3}$ is a 3×3 identity matrix, $\mathbf{u}$ is the vector of random additive genomic effects with design matrix $\mathbf{Z_1}$ (relating trait values to clones), $\mathbf{w}$ is the vector of random clone-by-location interaction effects with design matrix $\mathbf{Z_2}$ (relating trait values to clones-location combinations). For the c^{th} location, a column of $\mathbf{Z_c}$ may be all zeros if the clone represented by the column was not evaluated in that location. $\mathbf{K}$ is the additive genomic relationship matrix generated from SNP markers as above. In this model, the genomic value of a clone for the c^{th} location was es-

timated as $\hat{\mathbf{u}} + \hat{\mathbf{w}}_\mathbf{c}$. A more complete accounting of error terms would have included clone-by-year and clone-by-location-by-year terms in the model. While such a model would have characterized error in more detail we believe that its improvement of within-location estimation would have been marginal. Model 2.5 implies a compound symmetric structure [72] described below:

2.4.10 The uE model defined as a compound symmetry (CS) covariance structure model:

We define a model with compound symmetry covariance structure which is equivalent to the uE model (using same symbolism as defined in the uE model) as :

$$
\begin{aligned}
\mathbf{y} &= \mathbf{X}\beta + \mathbf{Z_2}\mathbf{w} + \mathbf{e} \\
\mathbf{y} = (\mathbf{y}_a^{'}, \mathbf{y}_b^{'}, \mathbf{y}_c^{'})^{'} \quad \mathbf{e} &= (\mathbf{e}_a^{'}, \mathbf{e}_b^{'}, \mathbf{e}_c^{'})^{'} \\
\mathbf{w} &\sim N\left(\mathbf{0}, (\sigma_u^2 + \sigma_w^2)\Phi \otimes K\right); \\
\Phi &= \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix} \\
\rho &= \frac{\sigma_u^2}{\sigma_u^2 + \sigma_w^2} \\
var(y) = \mathbf{V} &= (\sigma_u^2 + \sigma_w^2)(\mathbf{Z_2}(\mathbf{\Phi} \otimes \mathbf{K})\mathbf{Z_2^{'}}) + \sigma_e^2\mathbf{I} \\
\mathbf{Z_2} &= diag(\mathbf{Z_a}, \mathbf{Z_b}, \mathbf{Z_c})
\end{aligned}
\tag{2.6}
$$

The genomic effect from this CS model $\hat{\mathbf{w}}$ is equal to $\hat{\mathbf{u}} + \hat{\mathbf{w}}$ from the uE model.

The $\mathbf{Z_a}, \mathbf{Z_b}, \mathbf{Z_c}$ are design matrices relating records to clones in locations a, b and c respectively. Compared to the ME model described below which replaces $\mathbf{\Phi}$ with an unstructured covariance matrix with 9 parameters (6 for genetic and 3 for error (co)variances respectively), the CS model has 3 parameters σ^2_{u+w} (equivalent to $\sigma^2_u + \sigma^2_w$ in the uE model), σ^2_e and ρ. For any trait where the CS covariance structure best fits the data, it is expected that uE will provide more accurate GEBVs than the ME which will overfit the data. Furthermore, the uE defined here assumes a homogeneous variance across locations a, b and c. Although a CS model with heterogeneous variances can be fit, this was not the case for the uE model. This assumption will be incorrect if there are significant heterogeneous variances across these locations. In such a case, the ME model should provide more accurate breeding values.

2.4.11 The multivariate multi-environment (ME) model:

$$\begin{gathered} \mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e} \\ \mathbf{y} = (\mathbf{y}_a^{'}, \mathbf{y}_b^{'}, \mathbf{y}_c^{'})^{'} \quad u = (\mathbf{u}_a^{'}, \mathbf{u}_b^{'}, \mathbf{u}_c^{'})^{'}; \quad \mathbf{e} = (\mathbf{e}_a^{'}, \mathbf{e}_b^{'}, \mathbf{e}_c^{'})^{'} \end{gathered} \tag{2.7}$$

Where $\mathbf{y}$ is a vector of same trait in locations a, b, and c (corresponding to Ubiaja, Mokwa and Ibadan) recorded for n clones, $\mathbf{X}$ and $\mathbf{Z}$ design matrices are block diagonal matrices represented as $diag(\mathbf{X_a}, \mathbf{X_b}, \mathbf{X_c})$ and $diag(\mathbf{Z_a}, \mathbf{Z_b}, \mathbf{Z_c})$ respectively allowing for missing clones and observations. $\mathbf{X}$ is a design matrix for fixed effects β (with components as in model 2) and $\mathbf{Z}$ is a design matrix for random genomic effects $\mathbf{u}$. Following a multivariate normal distribution (Nm), the marginal density of $\mathbf{y}$ is given as:

$$\begin{gathered} (\mathbf{y}\,|\beta, \mathbf{R}, \mathbf{G}) \sim N_m(\mathbf{X}\beta, \mathbf{V}) \\ var(\mathbf{y}) = \mathbf{V} = \mathbf{Z}(\mathbf{G} \otimes \mathbf{K})\mathbf{Z}^T + \mathbf{R} \otimes \mathbf{I}); \quad \hat{\mathbf{u}} = (\mathbf{G} \otimes \mathbf{K})\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta) \end{gathered} \tag{2.8}$$

Given that d is number of locations being analyzed, **G** is a $d \times d$ symmetric and unstructured genomic covariance matrix while **R** is a d-dimensional diagonal error covariance matrix, **K** remains the additive genomic relationship matrix for n clones generated from SNP markers as above, **I** is an identity matrix and $\hat{\mathbf{u}}$ are the genomic estimated breeding values (GEBVs) of the clones and for the traits in the analysis. In this model, the error covariance matrix **R** is diagonal thus allowing heterogeneous variances of a trait for different locations but the covariances are fixed to zero following the assumption that no mechanism generates error covariances between a trait at multiple locations.

Estimation of the parameters in models (2,3,5 and 6) were performed using the average information (AI) REML procedure implemented in the airemlf90 program [62] from which BLUEs of fixed effects and BLUPs of random effects were obtained by solving the mixed model equations (MME) [5,6]. Custom R-scripts were used for cross validation.

2.4.12 Comparison of prediction accuracies:

We used a 5-fold cross validation scheme with 10 repeats for comparisons between the univariate and multivariate models. The same folds were used for the models in each scenario. We hereafter refer to predicted BLUPs or genomic effects from these models as Genomic EBVs (GEBVs). Prediction accuracies were calculated as a correlation of the validation fold GEBVs to their corresponding EGVs.

2.5 Results:

2.5.1 Scenario 1: MT vs uT models:

In Scenario 1, we observed that the prediction accuracies of the MT model were higher than those from the uT models for all traits and locations in our analysis (Table 2.2). On average (across traits and locations), the MT model had 59% higher prediction accuracy for VIGOR, 43% for RTNO, 27% for DM, 40% for MCMDS, 55% for FYLD and 18% for MCGM compared to the uT model. Averaged across traits and locations, the MT models were 40% more accurate than the uT models.

2.5.2 Scenario 2: ME vs uE models:

In Scenario 2, we observed different patterns of prediction accuracies of the uE and ME models. The ME model had higher prediction accuracies for DM and MCMDS at all locations. On average (across locations), the uE model had 2% better predictive ability for VIGOR and 1% for RTNO while the ME model had 32% better predictive ability for DM, 24% for MCMDS, 5% for FYLD, and 4% for MCGM. The ME model had 12% higher predictive ability than the uE model averaged across all traits and locations in the model. Trait correlations from the ME model representing the expected correlated responses to selection ranged from 0.21 to 0.66 for VIGOR, 0.36 to 0.54 for RTNO, 0.57 to 0.81 for DM, 0.68 to 0.87 for MCMDS, 0.31 to 0.52 for FYLD and 0.24 to 0.53 for MCGM. Thus, genetic effects for MCMDS and DM were more consistent across locations than were the genetic effects for other traits.

Table 2.2: **Cross validation prediction accuracies for GS models in scenarios 1 and 2.**

GS Scenario 1						
Single trait single environment (uT)				Multi-trait (MT)		
	Ubiaja	Mokwa	Ibadan	Ubiaja	Mokwa	Ibadan
VIGOR	0.24 (0.02)	0.16 (0.03)	0.42 (0.02)	0.41 (0.02)	0.31 (0.03)	0.58 (0.01)
RTNO	0.32 (0.02)	0.17 (0.02)	0.37 (0.02)	0.46 (0.02)	0.24 (0.03)	0.53 (0.02)
DM	0.60 (0.01)	0.33 (0.02)	0.51 (0.01)	0.72 (0.01)	0.46 (0.02)	0.64 (0.02)
MCMDS	0.49 (0.01)	0.37 (0.03)	0.59 (0.01)	0.69 (0.02)	0.60 (0.04)	0.74 (0.01)
FYLD	0.41 (0.02)	0.11 (0.03)	0.40 (0.01)	0.58 (0.02)	0.30 (0.03)	0.55 (0.02)
MCGM	0.38 (0.01)	0.50 (0.02)	0.58 (0.01)	0.48 (0.01)	0.56 (0.02)	0.69 (0.01)
GS Scenario 2						
Single trait multi-environment (uE)				Multi-environment (ME)		
	Ubiaja	Mokwa	Ibadan	Ubiaja	Mokwa	Ibadan
VIGOR	0.22 (0.01)	0.10 (0.01)	0.37 (0.01)	0.24 (0.01)	0.12 (0.02)	0.32 (0.01)
RTNO	0.29 (0.01)	0.11 (0.01)	0.34 (0.01)	0.27 (0.02)	0.13 (0.01)	0.34 (0.02)
DM	0.49 (0.01)	0.20 (0.02)	0.40 (0.01)	0.60 (0.01)	0.35 (0.01)	0.50 (0.01)
MCMDS	0.40 (0.01)	0.23 (0.01)	0.53 (0.01)	0.48 (0.01)	0.39 (0.02)	0.57 (0.01)
FYLD	0.38 (0.01)	0.10 (0.02)	0.35 (0.01)	0.37 (0.01)	0.12 (0.03)	0.36 (0.02)
MCGM	0.31 (0.01)	0.48 (0.01)	0.56 (0.01)	0.38 (0.02)	0.47 (0.01)	0.55 (0.01)

Prediction accuracies for MT and uT models (GS scenario 1) and for ME and uE models (GS scenario 2). The numbers in braces are standard deviations for cross validation repeat cycles.

2.6 Discussion:

2.6.1 Scenario 1: MT vs uT model:

Some studies have reported comparisons between MT and uT genomic selection models in simulation studies or also real data sets [57-59]. Guo et al., 2014 [59] and Calus et al., 2011 [58] in their studies with simulated data sets reported similar accuracies with small differences between their MT and uT models where accuracies for the MT models for low heritability traits were slightly higher when genomic correlations between the traits increased. VanRaden et al., 2014 [57] in research on Holstein and Jersey breed datasets from the US Dairy National evaluation program also reported similar accuracies with small differences between their MT and uT models for all the traits in their analysis. In several traits, their uT model accuracies were slightly higher than those of their MT model. Accuracies from the MT model may not be clearly better than those from the uT model for traits with high heritability, especially if these traits have complete phenotypic data are available [59]. Improvement in prediction accuracies for the MT model is accrued mostly for low heritability traits analyzed jointly with high heritability traits that have medium to high genomic correlations and low residual correlations [58, 59]. Our results were consistent with other studies [58,59] where our MT model had higher accuracies for most traits and locations in our analysis as a result of joint analysis of low heritability traits with other traits of higher heritabilities. Most of the genetic correlations between traits at all locations in the MT models were significant (substantial) with low error correlations (not shown). These significant genetic correlations contributed to the increased prediction accuracies observed for MT models compared to those of

uT models. Substantial increases in prediction accuracies of MT models were observed for VIGOR, RTNO, and FYLD which had mostly moderate to high genetic correlations with other traits at all locations even though their heritabilities were mostly low. We also observed (Table 2.6) that a combination of high genetic correlations and higher differences between the genetic variances of a trait across locations resulted in increased average prediction accuracies from the MT model. VIGOR, FYLD, RTNO and MCMDS benefited more from this combination while MCGM and DM benefited less.

For parental selections in specific locations, we recommend the use of MT models. Further studies on the selection gains based on these models are recommended to confirm this recommendation.

2.6.2 Scenario 2: ME vs uE model:

Again, some comparisons of different ME and uE genomic selection models have been done in plant breeding literature [67-69]. However, Burgueno et al., 2012 [35] conducted extensive modeling for multi-environment trials using pedigree and genomic markers and incorporated many covariance structures including diagonal, factor analytic (FA), identity and unstructured covariances for both the genomic and error components in their models. They observed higher prediction accuracies for their genomic uE model with a heterogeneous genomic variances and error variances (MED-D) compared to their genomic ME model with a FA genomic covariance structure and diagonal error covariance (MEFA-D) for most of the locations in their analysis based on their cross-validation scheme (CV1) [35]. This MED-D is a univariate model

with fewer parameters but may be compared to our uE model. Although the uE model assumed same genomic and error variances for all locations analyzed, total phenotypic variance was partitioned into direct clonal genomic, clone-by-location interaction and error variance components. Hence effects due to clones and clone-by-location interaction were combined to generate location specific GEBVs which may be compared to location specific GEBVs obtained in the MED-D model. Our results were in line with this study for the traits VIGOR and RTNO at all our locations where the uE model had higher prediction accuracies than the ME models and differed from this study for the traits DM and MCMDS at all locations where the ME models had higher prediction accuracies. However on average across locations and traits, the ME models had better predictability. To further understand the strength of the impact of GxE on the cassava core traits analyzed in this study, we utilized information from the proportion of total variation explained by clonal and clone-by-location effects from the uE model (Table 2.5). These reflected the fraction of total variance that was captured by whole genome GBS markers for these effects. From the total variation explained by markers, the effect of clone-by-location interaction were approximately 30% for VIGOR, 48% for RTNO, 12% for DM, 15% for MCMDS, 56% for FYLD and 46% for MCGM. This portrays the impact of location and hence strong clone-by-location interactions for the traits FYLD, RTNO, MCGM and VIGOR while being weak for the traits DM and MCGM. The ME models provided higher predictability and more accurate breeding values for the two traits DM and MCMDS.

The genetic correlations between the 3 locations for DM and MCMDS were relatively high ranging from 57 to 81% and 68 to 87% respectively (Table 2.4). These high correlations revealed that cassava DM and MCMDS were highly re-

peatable across the locations in our study suggesting that genotypes selected for these traits will perform comparably across the locations. From the genetic correlations in Table 2.4, improvement for RTNO and FYLD at Ubiaja will result in a correlated response of about 50% for these traits at Mokwa and about 35% at Ibadan. The low predicted correlated responses confirm that the environment had higher impacts on RTNO, FYLD, VIGOR and MCGM making their improvement more challenging. This makes a case for decentralized breeding especially for yield component traits. Breeding for good varieties that combine these core traits may be targeted towards specific locations or groups of locations with specific genotypes selected for these locations.

The ME model exploits the positive genomic correlations captured in its G matrix for prediction. The differences between the prediction accuracies of the ME and uE models were mainly due to the estimation of genetic covariances since their genomic variances were very similar. Genomic covariances from ME models are a reflection of GxE interactions of the trait of interest and ME breeding values capture both additive genotypic and additive genotype-by-environment effects. However lack of information from between-trait correlations (which are captured by MT models) in ME breeding values presents a challenge when selection decisions based on information from interconnection of multiple trait and multiple location data is desired. There is need for this information interconnection since valuable single environment and METs data are available.

Another potential use of ME models is for clustering of environments into target population of environments (TPEs). Using genomic correlations from these models, if correlated responses to selection of target traits are high for cer-

tain locations, then the locations can be grouped into a TPE. Regional breeding can commence within this TPE and all multi-location trials carried out within this TPE. As an example, Ubiaja and Ibadan can belong to same TPE considering the traits VIGOR, DM and MCMDS with correlated responses to selection ranging from 66 to 87% (Table 2.4).

Lower average (across locations) genetic correlations from the ME model for DM and MCMDS compared to those from the uE model (Table 2.7) resulted in 32% and 24% increased prediction accuracies respectively for these traits. This implied that the unstructured covariance structure from the ME model provided a better fit for the DM and MCMDS METs data. However no significant differences were observed in the prediction accuracies of the uE and ME models for the traits VIGOR, RTNO, FYLD and MCGM even though the estimated genetic correlations from the uE model were lower for most of these traits (Table 2.7). This technically implies that the uE model with less parameters is more parsimonious and should be favored more than the ME for the traits VIGOR, RTNO, FYLD and MCGM. However, more research is needed to understand the impacts of ranking of clones based on GEBVs from the uE and ME models and how these rankings affects gains since their accuracies were not significantly different.

2.6.3 Parameter estimates and implications for cassava breeding:

The estimates of genomic correlations and heritabilities shown in Table 2.3 have interesting implications for cassava genetic improvement. MT model genomic

correlation estimates between RTNO and FYLD were high and positive at all locations (ranging from 0.65 to 0.8); those between RTNO and DM were neutral to positive (ranging from -0.003 to 0.20) while those between FYLD and DM ranged between -0.02 and 0.11. The genomic correlations between these core production traits (DM, RTNO and FYLD) indicate that a concurrent improvement of these traits is achievable. However, more replication in trials targeting these production traits will help reduce error variances and improve the accuracy of parental selections given the low heritabilities for FYLD and RTNO. VIGOR can also be improved concurrently with these production traits as it is mostly positively correlated with them (Table 2.3). The disease trait (MCMDS) showed moderate to strong negative genomic correlations with VIGOR and the production traits, which is favorable for cassava breeding in Africa especially where the cassava mosaic disease (CMD) pressure is high. Consequently, cassava breeders have tried to fix genes for CMD resistance [64, 65]. With the favorable genomic correlations between these target traits in mind, the merit index from MT breeding values should be efficient as it takes into account genomic correlations.

We would like to make it clear here that fitting MT and ME models is computationally expensive requiring in our case the estimation of 90 and 36 additional covariance parameters for MT and ME models respectively compared to the uT and uE models. We had a few thousand records to accurately estimate these parameters as the standard error of these estimates show in Tables 2.3 and 2.4. When these correlations are not significant, breeding values from univariate models suffice because MT models are not expected to result in improved prediction accuracies [66].

However the traits VIGOR, RTNO, FYLD and MCGM (with heritability less or equal to 0.3) will benefit more from the MT model than the uT model. This is due to the joint analysis of these traits with DM and MCMDS with higher heritabilities exceeding 0.4. These benefits were accrued due to significant genetic correlations (exceeding 0.1) of these low heritability traits with other traits of high heritability. The significant genetic correlations between the traits analyzed in this study explain the higher accuracies observed for the MT model compared to the uT at all locations in this study (Table 2.2).

2.7 Conclusion:

The effectiveness of a breeding program is evaluated by its ability to provide adapted and productive varieties to the farming community in the TPEs it serves. To achieve this goal for the cassava breeding program at IITA, we recommend a decentralized breeding strategy for the different agroecological zones in Nigeria using total merit indices based on MT breeding values. Further studies should be conducted to understand how much selection gain can be made using this strategy. ME models provided less of an improvement in prediction accuracy but were useful for understanding GxE.

Table 2.3: **Genetic correlations and heritabilities for analyzed traits.** Plot-basis heritabilities on diagonal, genetic correlations from the MT model off diagonal and standard errors in braces.

Ubiaja						
	VIGOR	RTNO	DM	MCMDS	FYLD	MCGM
VIGOR	**0.16**					
RTNO	0.63 (0.007)	**0.21**				
DM	0.27 (0.014)	0.19 (0.009)	**0.42**			
MCMDS	-0.67 (0.020)	-0.53 (0.013)	-0.22 (0.025)	**0.62**		
FYLD	0.62 (0.009)	0.80 (0.008)	0.11 (0.012)	-0.42 (0.017)	**0.26**	
MCGM	0.05 (0.006)	-0.03 (0.004)	-0.17 (0.009)	0.22 (0.012)	-0.08 (0.006)	**0.1**
Mokwa						
	VIGOR	RTNO	DM	MCMDS	FYLD	MCGM
VIGOR	**0.06**					
RTNO	-0.11 (0.008)	**0.16**				
DM	0.12 (0.015)	-0.003 (0.010)	**0.31**			
MCMDS	-0.03 (0.016)	-0.35 (0.011)	-0.14 (0.020)	**0.64**		
FYLD	0.04 (0.010)	0.65 (0.008)	-0.15 (0.013)	-0.18 (0.013)	**0.21**	
MCGM	0.32 (0.008)	-0.15 (0.006)	-0.02 (0.011)	-0.03 (0.011)	-0.1(0.007)	**0.26**
Ibadan						
	VIGOR	RTNO	DM	MCMDS	FYLD	MCGM
VIGOR	**0.19**					
RTNO	0.46 (0.014)	**0.26**				
DM	0.18 (0.016)	0.20 (0.015)	**0.37**			
MCMDS	-0.64 (0.033)	-0.52 (0.029)	-0.13 (0.032)	**0.77**		
FYLD	0.34 (0.019)	0.77 (0.020)	-0.02 (0.019)	-0.44 (0.037)	**0.35**	
MCGM	-0.14 (0.012)	0.16 (0.011)	-0.08 (0.013)	0.11 (0.026)	0.11 (0.015)	**0.22**

Table 2.4: **Genetic correlations from the multi-environment analysis.** Genetic correlation estimates from the ME model are shown with the standard error of estimates in braces.

		Ubiaja	Mokwa
VIGOR	**Mokwa**	0.39 (0.023)	
	Ibadan	0.66 (0.027)	0.21 (0.033)
RTNO	**Mokwa**	0.54 (0.076)	
	Ibadan	0.36 (0.074)	0.38 (0.080)
DM	**Mokwa**	0.57 (0.004)	
	Ibadan	0.81 (0.002)	0.77 (0.001)
MCMDS	**Mokwa**	0.80 (0.031)	
	Ibadan	0.87 (0.048)	0.68 (0.035)
FYLD	**Mokwa**	0.52 (0.020)	
	Ibadan	0.31 (0.024)	0.33 (0.041)
MCGM	**Mokwa**	0.34 (0.006)	
	Ibadan	0.24 (0.008)	0.53 (0.010)

Table 2.5: **Proportion of explained variance by clonal and clone-by-location effects based on whole genome markers from the uE model.**

	Variance explained by effect (%)	
TRAIT	**Clone x Location**	**Clone**
VIGOR	4.52	10.70
RTNO	8.70	9.54
DM	4.59	33.52
MCMDS	10.00	58.90
FYLD	13.01	10.11
MCGM	7.00	8.33

Table 2.6: **Range of genetic correlations, genetic variances and the percentage increase in prediction accuracy from the MT model.** The σ_u^2 were estimated genetic variances for Ubiaja, Mokwa or Ibadan from the MT model for six target traits. ρ_{max} and ρ_{min} were maximum and minimum genetic correlations between a target trait and other traits from the MT model at all three locations.

TRAIT	$\|\rho_{\mathbf{max}}\| - \|\rho_{\mathbf{min}}\|$	$\frac{\mathbf{max}(\sigma_{\mathbf{u}}^2)-\mathbf{min}(\sigma_{\mathbf{u}}^2)}{\mathbf{max}(\sigma_{\mathbf{u}}^2)+\mathbf{min}(\sigma_{\mathbf{u}}^2)}$	**% Increase in accuracy across locations**
VIGOR	0.64	0.63	59
RTNO	0.80	0.45	43
DM	0.27	0.18	27
MCMDS	0.64	0.57	40
FYLD	0.78	0.50	55
MCGM	0.30	0.57	18

Table 2.7: **Estimated genetic correlations from the ME and uE models for six cassava traits.** ρ_{uE}, the genetic correlation from the CS model was estimated using variance components from the uE model while ρ_{ME} were genetic correlations from the ME model. $\bar{\rho}_{ME}$ represents mean of ME genetic correlations across locations while % accuracy increase reflects increased ME model accuracies over those of the uE across all locations.

TRAIT	$\rho_{\mathbf{uE}} = (\frac{\sigma^2_{\mathbf{u}}}{\sigma^2_{\mathbf{u}}+\sigma^2_{\mathbf{w}}})$	$\bar{\rho}_{\mathbf{ME}}$	$\mathbf{max}(\rho_{\mathbf{ME}}) - \mathbf{min}(\rho_{\mathbf{ME}})$	$\frac{\mathbf{max}(\sigma^2_{\mathbf{u}})-\mathbf{min}(\sigma^2_{\mathbf{u}})}{\mathbf{max}(\sigma^2_{\mathbf{u}})+\mathbf{min}(\sigma^2_{\mathbf{u}})}$	**% accuracy increase**
VIGOR	0.65	0.42	0.45	0.33	-2
RTNO	0.36	0.43	0.18	0.36	-1
DM	0.79	0.72	0.24	0.10	32
MCMDS	0.82	0.78	0.19	0.47	24
FYLD	0.29	0.39	0.21	0.31	5
MCGM	0.35	0.37	0.29	0.43	4

2.8 References:

[1] The International Plant Names Index (IPNI). http://www.ipni.org/ipni/idPlantNameSearch.do?id=351790-1. Accessed October 31, 2015.

[2] Taylor, N., Chavarriaga, P., Raemakers, K., Siritunga, D., & Zhang, P. (2004). Development and application of transgenic technologies in cassava. Plant Molecular Biology, 56(4), 671-688.

[3] Moorthy, S. N. (2002). Physicochemical and functional properties of tropical tuber starches: a review. StarchStrke, 54(12), 559-592.

[4] Balagopalan, C. (2002). Cassava utilization in food, feed and industry. Cassava: Biology, production and utilization, 301-318.

[5] Henderson, C.R., 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics, pp.423-447.

[6] Henderson, C. R. "Estimation of variances and covariances under multiple trait models." Journal of Dairy Science 67, no. 7 (1984): 1581-1589.

[7] Smith, H.F. 1936. A discriminant function of plant selection. Ann. Eugenics, 7: 240-250.

[8] Hazel, Lanoy Nelson. "The genetic basis for constructing selection indexes." Genetics 28, no. 6 (1943): 476-490.

[9] Ducrocq, V.; Wiggans, G.; Garrick, D. J.; Ruvinsky, A.; CABI, Wallingford, UK, The genetics of cattle, 2015, Ed. 2, pp 380-384.

[10] Meuwissen T.H.E., Hayes B.J., Goddard M.E. (2001) Prediction of total genetic value using genome wide dense marker maps. Genetics, 157, 18191829.

[11] Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. "Invited review: Genomic selection in dairy cattle: Progress and challenges." Journal of dairy science 92, no. 2 (2009): 433-443.

[12] Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, and S.E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6:e19379. doi:10.1371/journal.pone.0019379

[13] VanRaden, P.M. and Wiggans, G.R. (1991). Derivation, calculation, and use of national animal model information. Journal of Dairy Science 74, 2737-2746.

[14] Goddard, Mike. "Genomic selection: prediction of accuracy and maximisation of long term response." Genetica 136.2 (2009): 245-257.

[15] Van der Werf, J., Van Arendonk, J.A.M., de Vries, A.G., 1992. Improving selection of pigs using correlated characters. 43rd Ann. EAAP meeting, selection of pigs using correlated characters. 43rd Ann. EAAP meeting, Madrid, Spain.

[16] Ducrocq, V., 1994. Multiple trait prediction: principles and problems. in Proc. 5th World Congr. Genet. App. Livest. Prod., Guelph, Ontario, Canada. Vol. 18, 455-462.

[17] Colleau, Jean-Jacques, Vincent Ducrocq, Didier Boichard, and Hlne Larroque. "Approximate multi-trait BLUP evaluation to combine functional

traits information." Interbull Bulletin 23 (1999): 151.

[18] Schaeffer, L. R. "Multiple-country comparison of dairy sires." Journal of Dairy Science 77, no. 9 (1994): 2671-2678.

[19] Schaeffer, L. R. "Multiple trait international bull comparisons." Livestock Production Science 69, no. 2 (2001): 145-153.

[20] Thompson, R., and K. Meyer. "A review of theoretical aspects in the estimation of breeding values for multi-trait selection." Livestock Production Science 15, no. 4 (1986): 299-313.

[21] Ducrocq V, Boichard D, Barbat A, Larroque H. Implementation of an approximate multi-trait BLUP evaluation to combine production traits and functional traits into a total merit index. In Proceedings of the 52nd Annual Meeting of the European Association for Animal Production: 26-29 August 2001; Budapest; 2001.

[22] Lassen J, Srensen MK, Madsen P, Ducrocq V. A stochastic simulation study on validation of an approximate multitrait model using preadjusted data for prediction of breeding values. J Dairy Sci. 2007;90:300211.

[23] Lassen J, Srensen MK, Madsen P, Ducrocq V. An approximate multi-trait model for genetic evaluation in dairy cattle with a robust estimation of genetic trends. Genet Sel Evol. 2007;39:35367.

[24] Segelke, D., Reinhardt, F., Liu, Z. and Thaller, G., 2014. Prediction of expected genetic variation within groups of offspring for innovative mating schemes.Genetics Selection Evolution, 46(1), p.1.

[25] Habier, D., Fernando, R. L., & Dekkers, J. C. (2008, 02). The impact of genetic relationship information on genome-assisted breeding values. Genetics. doi:10.1534/genetics.107.081190

[26] Malosetti, M., Ribaut, J.M. and van Eeuwijk, F.A., 2014. I. 6 The statistical analysis of multienvironment data: modelling genotype-by-environment interaction and its genetic basis. Drought phenotyping in crops: From theory to practice, 4(44), p.53.

[27] Halsey, M.E., Olsen, K.M., Taylor, N.J. and Chavarriaga-Aguirre, P., 2008. Reproductive Biology of Cassava (Crantz) and Isolation of Experimental Field Trials. Crop science, 48(1), pp.49-58.

[28] Kawano, K., Amaya, A., Daza, P. and Rios, M., 1978. Factors affecting efficiency of hybridization and selection in cassava. Crop Science, 18(3), pp.373-376.

[29] Kawano, K. 1980. Cassava. p. 225233. In W.R. Fehr and H.H. Hadley (ed.) Hybridization of crop plants. ASA and CSSA, Madison, WI.

[30] Smith, A.B., B.R. Cullis, and R. Thompson. 2001. Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. Biometrics 57:11381147. doi:10.1111/j.0006-341X.2001.01138.x

[31] Falconer, D.S. 1952. The problem of environment and selection. Am. Nat. 86:293-298

[32] Crossa, J., J. Burgueo, P.L. Cornelius, G. McLaren, R. Trethowan, and A. Krishnamachari. 2006. Modeling genotype environment interaction using additive genetic covariances of relatives for predicting breeding values of

wheat genotypes. Crop Sci.46:17221733. doi:10.2135/cropsci 2005.11-0427

[33] Burgueo, J., J. Crossa, P.L. Cornelius, and R.-C. Yang. 2008. Using factor analytic models for joining environments and genotypes without crossover genotype environment interaction. Crop Sci. 48:12911305. doi:10.2135/cropsci 2007.11.0632

[34] Burgueo, J., J. Crossa, J. Miguel Cotes, F. San Vicente, and B. Das. 2011. Prediction assessment of linear mixed models for multienvironment trials. Crop Sci. 51:944954. doi:10.2135/cropsci 2010.07.0403

[35] Burgueo, J., de los Campos, G., Weigel, K., & Crossa, J. (2012). Genomic prediction of breeding values when modeling genotype environment interaction using pedigree and dense molecular markers. Crop Science, 52(2), 707-719.

[36] de los Campos, G., and D. Gianola. 2007. Factor analysis models for structuring covariance matrices of additive genetic effects: A Bayesian implementation. Genet. Sel. Evol. 39:481494. doi:10.1186/1297-9686-39-5-481

[37] Meyer, Karin. "Factor-analytic models for genotype x environment type problems and structured covariance matrices." Genet Sel Evol 41, no. 11 (2009).

[38] Tsuruta, S., Lourenco, D. A. L., Misztal, I., and Lawlor, T. J. (2015). Genotype by environment interactions on culling rates and 305-day milk yield of Holstein cows in 3 US regions.Journal of Dairy Science.

[39] Kolmodin, R., F. Strandberg, P. Madsen, J. Jensen, and H. Jorji-ani. 2002. Genotype by environment interaction in Nordic dairy Acta Agric. Scand.

Anim. cattle studied using reaction norms. Sci. 52:1124.

[40] Windig, J. J., M. P. L. Calus, and R. F. Veerkamp. 2005. Influence of herd environment on health and fertility and their relationship with milk production. J. Dairy Sci. 88:335347.

[41] Cooper, M., and I.H. DeLacy. 1994. Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. Theor. Appl. Genet. 88(5): 561572.

[42] Malosetti, Marcos, Jean Marcel Ribaut, Mateo Vargas, Jos Crossa, and Fred A. Van Eeuwijk. "A multi-trait multi-environment QTL mixed model with an application to drought and nitrogen stress trials in maize (Zea mays L.)." Euphytica 161, no. 1-2 (2008): 241-257.

[43] Alimi, N. A., M. C. A. M. Bink, J. A. Dieleman, J. J. Magn, A. M. Wubs, A. Palloix, and F. A. van Eeuwijk. "Multi-trait and multi-environment QTL analyses of yield and a set of physiological traits in pepper." Theoretical and applied genetics 126, no. 10 (2013): 2597-2625.

[44] Garrick, D. J., Taylor, J. F., and Fernando, R. L. (2009). Deregressing estimated breeding values and weighting information for genomic regression analyses.Genet Sel Evol, 41(55), 10-1186.

[45] Ostersen, Tage, Ole F. Christensen, Mark Henryon, Bjarne Nielsen, Guosheng Su, and Per Madsen. "Deregressed EBV as the response variable yield more reliable genomic predictions than traditional EBV in pure-bred pigs." Genetics Selection Evolution 43, no. 1 (2011): 1-6.

[46] Zhou, Xiang, and Matthew Stephens. "Efficient multivariate linear mixed model algorithms for genome-wide association studies." Nature methods 11, no. 4 (2014): 407-409.

[47] Deniz Akdemir and Okeke Uche Godfrey (2014). EMMREML: Fitting mixed models with known covariance structures. R package version 2.0. http://CRAN.R-project.org/package=EMMREML

[48] Maziya-Dixon, B., A.G.O. Dixon, and A.-R.A. Adebowale. 2007. Targeting different end uses of cassava: Genotypic variations for cyanogenic potentials and pasting properties. Int. J. Food Sci. Technol. 42:969976. doi:10.1111/j.1365-2621.2006.01319.x

[49] Okechukwu, R.U., and A.G.O. Dixon. 2008. Genetic gains from 30 years of cassava breeding in Nigeria for storage root yield and disease resistance in elite cassava genotypes. J. Crop Improv. 22:181208. doi:10.1080/15427520802212506

[50] Ly, Delphine, et al. "Relatedness and genotype environment interaction affect prediction accuracies in genomic selection: A study in cassava." Crop Science53.4 (2013): 1312-1325.

[51] Glaubitz, J., T. Casstevens, R. Elshire, J. Harriman, and E.S. Buckler. 2012. TASSEL 3.0 genotyping by sequencing (GBS) pipeline documentation. Edward S. Buckler, USDA-ARS, Ithaca, NY. `http://www.maizegenetics.net/tassel/docs/TasselPipelineGBS.pdf` (accessed 3. Jan. 2014)

[52] Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., ... and Rokhsar, D. S. (2012). Phytozome: a comparative platform

for green plant genomics. Nucleic acids research, 40(D1), D1178-D1186. (http://www.phytozome.net; accessed 1 July, 2014).

[53] Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. "Regularization paths for generalized linear models via coordinate descent." Journal of statistical software 33, no. 1 (2010): 1.

[54] R Development Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org (January, 2014)

[55] Douglas Bates, Martin Maechler, Ben Bolker and Steven Walker (2013). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-5. http://CRAN.R-project.org/package=lme4.

[56] Pfeiffer, Christina, Birgit Fuerst-Waltl, Hermann Schwarzenbacher, Franz Steininger, and Christian Fuerst. "A comparison of methods to calculate a total merit index using stochastic simulation." Genetics Selection Evolution 47, no. 1 (2015): 36.

[57] VanRaden, P. M., Tooker, M. E., Wright, J. R., Sun, C., and Hutchison, J. L. (2014). Comparison of single-trait to multi-trait national evaluations for yield, health, and fertility. Journal of dairy science, 97(12), 7952-7962.

[58] Calus, M. P., and Veerkamp, R. F. (2011). Accuracy of multi-trait genomic selection using different methods. Genet Sel Evol, 43(1), 1-14.

[59] Guo, G., Zhao, F., Wang, Y., Zhang, Y., Du, L., and Su, G. (2014). Comparison of single-trait and multiple-trait genomic prediction models. BMC genetics, 15(1), 30.

[60] Ceballos, H., Kawuki, R.S., Gracen, V.E., Yencho, G.C. and Hershey, C.H., 2015. Conventional breeding, marker-assisted selection, genomic selection and inbreeding in clonally propagated crops: a case study for cassava. Theoretical and Applied Genetics, 128(9), pp.1647-1667.

[61] Wimmer, V., Albrecht, T., Auinger, H.J. and Schn, C.C., 2012. synbreed: a framework for the analysis of genomic prediction data using R. Bioinformatics,28(15), pp.2086-2087.

[62] Misztal, I., S. Tsuruta, T. Strabel, B. Auvray, T. Druet, and D. Lee. 2002. BLUPF90 and related programs (BGF90). Proc. 7th World Congr. Genet. Appl. Livest. Prod., Montpellier, France. CD-ROM Communication 28:07

[63] VanRaden, P.M., 2008. Efficient methods to compute genomic predictions.Journal of dairy science, 91(11), pp.4414-4423.

[64] Jennings, D.L., 1972. Breeding for resistance to African cassava mosaic disease: Progress and prospects.

[65] Wolfe, M.D., Rabbi, I.Y., Egesi, C., Hamblin, M., Kawuki, R., Kulakow, P., Lozano, R., Carpio, D.P.D., Ramu, P. and Jannink, J.L., 2016. Genome-Wide Association and Prediction Reveals Genetic Architecture of Cassava Mosaic Disease Resistance and Prospects for Rapid Genetic Improvement. The Plant Genome, 9(2).

[66] Schaeffer, L.R., 1984. Sire and cow evaluation under multiple trait models. Journal of Dairy Science, 67(7), pp.1567-1580.

[67] Oakey, H., Cullis, B., Thompson, R., Comadran, J., Halpin, C. and Waugh, R., 2016. Genomic selection in multi-environment crop trials. G3: Genes,

Genomes, Genetics, 6(5), pp.1313-1326.

[68] Cuevas, J., Crossa, J., Soberanis, V., Prez-Elizalde, S., Prez-Rodrguez, P., Campos, G.D.L., Montesinos-Lpez, O.A. and Burgueo, J., 2016. Genomic prediction of genotype environment interaction kernel regression models. The plant genome.

[69] Cuevas, J., Crossa, J., Montesinos-Lpez, O.A., Burgueo, J., Prez-Rodrguez, P. and de los Campos, G., 2017. Bayesian Genomic Prediction with Genotype Environment Interaction Kernel Models. G3: Genes, Genomes, Genetics, 7(1), pp.41-53.

[70] Rutkoski, J., Benson, J., Jia, Y., Brown-Guedira, G., Jannink, J.L. and Sorrells, M., 2012. Evaluation of genomic prediction methods for Fusarium head blight resistance in wheat. The Plant Genome, 5(2), pp.51-61.

[71] Wolfe, M.D., Kulakow, P., Rabbi, I.Y. and Jannink, J.L., 2016. Marker-based estimates reveal significant non-additive effects in clonally propagated cassava (Manihot esculenta): implications for the prediction of total genetic value and the selection of varieties. G3: Genes, Genomes, Genetics, pp.g3-116.

[72] Piepho, H.P. and Pillen, K., 2004. Mixed modelling for QTL environment interaction analysis. Euphytica, 137(1), p.147.

CHAPTER 3

REGIONAL HERITABILITY MAPPING PROVIDES INSIGHTS INTO DRY MATTER (DM) CONTENT IN AFRICAN WHITE AND YELLOW CASSAVA POPULATIONS.

3.1 Abstract:

The HarvestPlus program for cassava (Manihot esculenta Crantz) fortifies cassava with beta-carotene by breeding for carotene-rich tubers (yellow cassava). However, a negative correlation between yellowness and dry matter (DM) content has been identified. Here, we investigated the genetic control of DM in white and yellow cassava subpopulations. We used regional heritability mapping (RHM) to associate DM to genomic segments in both subpopulations. Significant segments were subjected to candidate gene analysis and we attempted to validate candidates using prediction accuracies. The RHM procedure was validated using a simulation approach. The RHM revealed significant hits for white cassava on chromosomes 1, 4, 5, 10, 17 and 18 while hits for the yellow were on chromosome 1. Candidate gene analysis revealed genes in the carbohydrate biosynthesis pathway including the plant serine-threonine protein kinases (SnRKs), UDP-glycosyltransferases, UDP-sugar transporters, invertases, pectinases, and some regulatory genes. Validation using 1252 unique identifiers from the SnRK gene family genome-wide recovered 50% of the predictive accuracy of whole genome SNPs for DM while validation using 53 likely (extracted from literature) genes from significant segments recovered 32%. Genes including an acid invertase, a neutral/alkaline invertase and a glucose-6-phosphate isomerase were validated based on an a priori list for the cassava starch path-

way and also a fructose-biphosphate aldolase from the calvin cycle pathway. The power of the RHM procedure was estimated at 47 percent when the causal QTL generated 10% of the phenotypic variance with sample size of 451. Cassava DM genetics is complex. RHM may be useful for complex traits.

3.2 Core ideas:

- Regional heritability mapping (RHM) is effective for understanding the genetic architecture of complex traits in cassava.
- Prediction accuracies can reflect the impact of genomic segments on cassava dry matter (DM) content.
- Serine-threonine protein kinases (SnRKs) are candidates positionally associated with cassava DM.
- Prediction accuracy of SnRKs for cassava DM was 50% of the total accuracy from genome-wide SNPs.

3.3 List of Abbreviations:

- Dry matter content (DM)
- Fresh root yield (FYLD)
- Linkage disequilibrium (LD)
- Quantitative trait loci (QTL)
- Regional heritability mapping (RHM)
- Serine-threonine protein kinases (SnRKs)

- Single nucleotide polymorphisms (SNPs)
- Genotype-by-sequencing (GBS)
- Genome-wide association analysis (GWAS)

3.4 Background:

Cassava currently ranks as the sixth world staple crop consumed by more than 500 million people in Africa, Asia and South America (El-Sharkawy, 2003). It was originally a perennial shrub but is cultivated now as an annual for its starchy root (El-Sharkawy, 2003). It is an outbreeding species and considered to be an amphidiploid or sequential allopolyploid (El-Sharkawy, 2003). The crop is clonally propagated by mature woody stem cuttings called stakes, which are 15-30 cm long and planted mostly inclined on ridged soils (Keating et al., 1988). Botanical seeds are used mainly in breeding programs with up to three seeds produced per pod (Iglesias et al., 1994, Iglesias and Hershey, 1994). Storage roots are generally harvested 7 to 24 months after planting (El-Sharkawy, 2003). Dry matter (DM) is the major product from cassava roots apart from moisture and traces of water-soluble vitamins and pigments (Holleman and Aten, 1956; Barrios and Bressani, 1967; Lim, 1968). On average, cassava DM is made up of about 90% carbohydrates (mainly starch), 2% protein, 1% fat, 3% minerals and ash and 4% fiber (Holleman and Aten, 1956; Barrios and Bressani, 1967; Lim, 1968). This starch deposit makes cassava attractive for the food industry and other industries that rely heavily on starch as their primary raw material (Lim, 1968). The value of cassava derives from a combination of fresh root yield and the percentage DM that can be extracted from fresh roots, referred to as

dry yield. Fresh cassava roots with high DM content are also preferred by local farmers and processors (Kawano et al., 1987; Safo-Kantanka and Owusu-Nipah, 1992; Enidiok et al., 2008) who transform cassava roots into valuable staples consumed by many in developing countries. With 263 million metric tons produced in 2012 (FAOSTAT Database, 2013), cassava has become an indispensable staple in the world and improvement of cassava for high dry yield is needed. This improvement should also endeavor to increase micronutrient content, as it is much needed in the cassava consuming regions of the world. Biofortification is a successful genetic improvement technique for increasing micronutrient content in staple crops (Meenakshi et al., 2010a; Bouis et al., 2011) and represents a promising approach for solving the problem of micronutrient malnutrition around the world (Meenakshi et al., 2010a; Meenakshi et al., 2010b; Pfeiffer and McClafferty, 2007).

The target of biofortification is to increase the content of essential micronutrients such as Iron, Zinc, and Vitamin A (Meenakshi et al., 2010a; Meenakshi et al., 2010b; Pfeiffer and McClafferty, 2007), hence improving the health of millions of people who depend on these staples for daily nutrition. The biofortification process is facilitated by plant breeding (Meenakshi et al., 2010a; Bouis et al., 2011). Since the early 2000s, the HarvestPlus initiative (Meenakshi et al., 2010b; Pfeiffer and McClafferty, 2007) has been tasked with biofortification of staple crops including cassava, sweet potato, maize, rice and wheat. Biofortification of cassava is geared towards breeding varieties containing increased levels of provitamin A, or beta-carotene, a precursor for vitamin A. The so-called yellow cassava (Liu et al., 2010; Plus, 2009; Aniedu and Omodamiro, 2012; La Frano et al., 2013) is designed to address public health issues including child mortality, impaired vision and night blindness, reduced immunity to diseases and other

consequences of vitamin A deficiency (Liu et al., 2010; Plus, 2009).

Breeding for required levels of provitamin A necessitates the accumulation of beta-carotene in cassava roots (Aniedu and Omodamiro, 2012; La Frano et al., 2013). Many breeding programs use yellow flesh color as a proxy for measuring beta-carotene amount in cassava despite the fact that yellowness is more of an indication of total carotenoids in the root (Chvez et al., 2005; Ssemakula et al., 2007; Akinwale et al., 2010). This protocol is used to visually pre-select lines containing beta-carotenoids prior to quantification of different carotenoid levels using HPLC protocols (Kimura et al., 2007; Adewusi and Bradbury, 1993). Breeding for farmer preferred bio-fortified cassava involves the development of high yielding clones with high DM and high beta-carotene accumulation in a single clone or variety (Ceballos et al., 2004; Raji et al., 2007). Incorporating all these characteristics in a single variety of cassava makes for a challenging breeding task. Some studies have shown that there is a negative genetic correlation between DM and yellow root flesh color in cassava making this breeding task even more challenging since the target is towards full adoption of pro-vitamin A varieties by local farmers and processors (Akinwale et al., 2010; Vimala et al., 2008). It is therefore useful to understand the genetic control of DM content and beta-carotene accumulation in cassava to facilitate the breeding of farmer-preferred varieties.

Regional heritability mapping (RHM) is a relatively new procedure for identifying loci affecting quantitative traits (Nagamine et al., 2012; Riggio and Pong-Wong, 2014; Riggio et al., 2013; Shirali et al., 2015). Unlike single marker GWAS methods which the lack power to detect rare genetic variants (Bodmer and Tomlinson, 2010; Gibson, 2012; Wood et al., 2014), RHM can capture both rare and

common genetic variants giving it more power to identify loci that cannot be detected by standard GWAS (Nagamine et al., 2012; Riggio and Pong-Wong, 2014; Riggio et al., 2013). The RHM has been shown to detect both common and rare genetic variants implicated in disease traits in human genomics (Shirali et al., 2015; Uemoto et al., 2013; Zeng et al., 2016) and recently in tree genomics (Resende et al., 2017). RHM is a suitable method for capturing the effect of a genomic block or segment since it can identify genomic segment-trait associations for regions spanning multiple loci (Nagamine et al., 2012; Riggio and Pong-Wong, 2014; Riggio et al., 2013; Caballero et al., 2015). A multi-marker mapping approach like the RHM may identify both common and rare variants involved in the expression of DM in white and yellow subpopulations of African cassava. To the best of our knowledge, this is the first attempt to use the RHM procedure in an annual crop.

The objectives of this study were:

1. To understand the genetic basis of DM in white and yellow root African cassava populations.
2. To determine the power of the RHM procedure to detect genomic segments carrying QTL using the hide-a-causal-SNP procedure.

3.5 Materials and Methods:

Cassava phenotypic data for discovery: We used phenotypic data collected from the Genetic Gain (GG) population trials conducted by the cassava breeding program at the Institute of Tropical Agriculture (IITA), Ibadan, Nigeria for our analysis. The GG population (713 clones) is an elite population bred from the 1970s to 2007 by the cassava breeding program at the IITA (Maziya-Dixon et al., 2007; Okechukwu and Dixon, 2008; Ly et al., 2013). Most GG clones are of African origin with very good performance such that they were advanced to advanced to multi-environment uniform yield trials. For this study, we used clonal evaluation trials (CETs) of the GG population planted in an augmented design. The CET uses an unreplicated incomplete block design consisting of a layout of between 18 to 30 blocks with 22 accessions and two checks in each block. Accession plots were a single row (1m x 1m spacing) of five-plant stands without borders. All checks were included in the analysis. A few trials were replicated twice. These trials were conducted in three locations in Nigeria: Ibadan (7.40 N, 3.90 E), Mokwa (9.3 N, 5.0 E), and Ubiaja (6.66 N, 6.38 E) between 2013 and 2015. Three core agronomic traits were measured for these trials including fresh weight of harvested roots expressed in tons per hectares (T/ha) (FYLD), percentage dry matter (DM) of storage roots, which measures root dry weight as the percentage of the root fresh weight, and pulp color (PLPCOL) a binary trait rated on a scale from 1 (white flesh to light cream root) to 2 (deep cream to yellow flesh root). The DM trait was measured using the oven method: 100g grated root sample (with thorough mixing of 10-15 randomly selected roots from a plot) were collected per accession and oven dried. DM content was then measured as residual weight after oven drying. We further divided the GG population (713

clones) into two subpopulations of white (451 clones) and yellow (262 clones) cassava using the PLPCOL trait where clones with a score of 1 for this trait were grouped into the white population and those with score 2 into the yellow population.

3.5.1 Cassava phenotypic data for validation:

To validate results from the RHM analysis, we used data from a population called the GS-C, which consisted of progenies of clones from the GG population described above. Phenotypes from the GS-C1 were obtained from clonal evaluation trials (CETs) of 1,651 clones split into trials at three locations: Ibadan, Mokwa and Ikenne (652N 343E). These trials were planted using an augmented design consisting of between 20 to 30 blocks with 22-24 clones and two checks in each block. Plots were a single row of five-plant stands (1m x 1m spacing) without borders and without replication and trials were planted during 2014 and 2015. Cassava trait measurements for this population were as described earlier, except that no strict distinction between yellow and white flesh color was used because the GS-C1 were majorly white and cream clones; thus we performed validation analysis using all clones.

3.5.2 Cassava genotype data:

DNA was extracted using DNeasy Plant Mini Kits (Qiagen) from 713 clones from the 2013 Genetic Gain trial at IITA and was quantified using PicoGreen. Genotyping-by-sequencing (GBS) was used for genotyping (Elshire et al., 2011)

these clones. Six 95-plex and one 75-plex ApeKI libraries were constructed and sequenced on Illumina HiSeq, one lane per library. Single nucleotide polymorphisms (SNPs) were called from the sequence data using the TASSEL pipeline version 4.0 (Glaubitz et al., 2012), using an alignment to the M. esculenta version-6 reference genome (Goodstein et al., 2012). The marker data was converted to dosage format (0, 1, 2) and missing genotypic data were imputed using the Beagle software (Ayres et al., 2011). The final data set consisted of 177,201 SNPs scored in 713 clones. Members of the GS-C1 used in the validation analysis were genotyped in 2014 as described above. SNPs from both populations were called together using the TASSEL pipeline (Glaubitz et al., 2012) and missing genotypes also imputed using Beagle (Ayres et al., 2011) yielding the same number of SNPs as above.

3.5.3 Data analysis:

Genome-wide Regional Heritability mapping (RHM):

RHM was carried out using the following procedure (Nagamine et al., 2012; Riggio and Pong-Wong, 2014; Riggio et al., 2013):

(a) Chromosomes were divided into 100 SNP segments in sliding windows with 50 SNPs overlapping between adjacent windows.

(b) A multikernel univariate mixed model was used to partition the genomic additive variation due to trait of interest into components of the target genomic segment and the whole genome SNP markers as follows:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\beta + \mathbf{Z}\mathbf{u_1} + \mathbf{Z}\mathbf{u_2} + \mathbf{e} \\ \mathbf{u_1} \sim N(\mathbf{0}, \sigma^2_{u_1} K_{u_1}); \quad \mathbf{u_2} &\sim N(\mathbf{0}, \sigma^2_{u_2} K_{u_2}); \quad \mathbf{e} \sim N(\mathbf{0}, \sigma^2_e I_{n\times n}) \\ var(y) = \mathbf{V} &= \mathbf{Z}(\sigma^2_{u_1}\mathbf{K}_{u_1})\mathbf{Z}^T + \mathbf{Z}(\sigma^2_{u_2}\mathbf{K}_{u_2})\mathbf{Z}^T + \sigma^2_e\mathbf{I} \\ \hat{\mathbf{u_1}} = (\sigma^2_{u_1}\mathbf{K}_{u_1})\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta}); \quad \hat{\mathbf{u_2}} &= (\sigma^2_{u_2}\mathbf{K}_{u_2})\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta}) \end{aligned} \tag{3.1}$$

Where $\mathbf{y}$ is a response variable (DM), $\mathbf{X}$ is a known incidence matrix for fixed effects β (including grand mean and a nested effect of Rep within Trial within Year within Location), $\mathbf{Z}$ is a known incidence matrix for clonal additive genomic effects $\mathbf{u_1}$ for the target genomic segment and u2 for the whole genome SNPs. $\mathbf{K_{u_1}}$ and $\mathbf{K_{u_2}}$ are the genomic relationship matrices calculated from the SNPs using the procedure of VanRaden (2008) as:

$$\mathbf{G} = \frac{\mathbf{M}\mathbf{M}^T}{2\Sigma p(1-p)} \tag{3.2}$$

where $\mathbf{G}$ is the genomic relationship matrix, $\mathbf{M}$ is a centered marker matrix coded as -1,0,1 and p is the major allele frequency vector. Other components of the model include the genomic variance for the target genomic segment $\sigma^2_{u_1}$ and the total genomic variance for the whole genome $\sigma^2_{u_1}$ is the genomic error variance and $\mathbf{e}$ are the residuals from the model. Model (4.1) was fit using the R EMMREML package (Akdemir and Okeke, 2014). Note that the $\mathbf{K_{u_2}}$ genomic relationship matrix serves to statistically control for population structure effects as the kinship matrix does in standard GWAS.

(c) Following model fit from step (b) above, genomic heritability for each target

genomic segment was computed as follows:

$$h^2 = \frac{\sigma^2_{u_1}}{\sigma^2_{u_1} + \sigma^2_{u_2} + \sigma^2_e} \tag{3.3}$$

where h^2 is genomic heritability for a target genomic segment and variance components are described above.

(d) A likelihood ratio test (LRT) was used to test the significance of target genomic segments with the alternative model as Model (4.1) and the null model as model (4.1) without the target genomic kernel component ie $\mathbf{y} = \mathbf{X}\beta + \mathbf{Zu_2} + \mathbf{e}$. This model was also fit using the EMMREML package (Akdemir and Okeke, 2014). P-values were obtained using the *pchisq* function in R (R Core Development Team, 2016).

(e) Local FDR (LFDR) was estimated using the R qvalue package (Sorey and Tibshirani, 2003; Storey et al., 2015).

(f) Genomic segment LFDRs were then plotted across the genome in a Manhattan plot with a cutoff of 0.05 used to assess significance.

We carried out the RHM procedure separately for the white and yellow cassava subpopulations of GG. No defined population structure was found on in the GG population in a previous GWAS study (Wolfe et al., 2015). Therefore, the genomic relationship matrix from the whole genome SNPs in the RHM was sufficient to account for structure in this analysis (in fact we refer to this more as background effect).

Candidate gene analysis:

We identified candidate genes from the significant hits of the RHM analysis based on annotations for the v6 M. esculenta genome on phytozome (Goodstein et al., 2012). We used plant physiology information to narrow down the list of genes associated with carbohydrate biosynthesis including genes functional in starch and sugar biosynthesis, cell wall loosening and degradation, and root sink and plant growth pathways. We carried out validation tests on selected candidates based on prediction accuracies on the GS-C1 population as described below.

Validation models and procedures:

We conducted validation analyses for the significant hits of the RHM analysis and for the RHM procedure itself. Validation here was geared towards understanding the prediction accuracies obtained from genes and gene families on RHM significant segments. Validation proceeded as follows:

Validation using SnRK genes (a candidate gene family):

To obtain genotypic data for this analysis, we searched the Phytozome M. esculenta v6.1 web portal (Goodstein et al., 2012) using the keyword serine threonine kinases to recover all its instances in the cassava genome, resulting in 2,408 hits. We filtered the resulting list to remove all hits not containing gene ontology or Eukaryotic Orthologous Groups function definitions for the keyword serine threonine kinase. We then manually added genes containing known serine threonine kinases that did not contain a function definition, for example the

SNF1 gene. We extracted all markers within 2.5 kb of the start and end of each gene model using the Bedtools intersect function (Quinlan and Hall, 2010) resulting in 7,203 unique SNPs. We refer to these SNPs as candidate SNPs below. For validation of these candidate SNPs on the GS-C1 data we fit the following model:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\beta + \mathbf{Zs} + \mathbf{Zg} + \mathbf{e} \\ \mathbf{s} \sim N(\mathbf{0}, \sigma_s^2 K_s); \quad \mathbf{g} &\sim N(\mathbf{0}, \sigma_g^2 K_g); \quad \mathbf{e} \sim N(\mathbf{0}, \sigma_e^2 I_{n\times n}) \\ var(y) = \mathbf{V} &= \mathbf{Z}(\sigma_s^2 \mathbf{K}_s)\mathbf{Z}^T + \mathbf{Z}(\sigma_g^2 \mathbf{K}_g)\mathbf{Z}^T + \sigma_e^2 \mathbf{I} \\ \hat{\mathbf{s}} = (\sigma_s^2 \mathbf{K}_s)\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta}); \quad \hat{\mathbf{g}} &= (\sigma_g^2 \mathbf{K}_g)\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta}) \end{aligned} \tag{3.4}$$

Where $\mathbf{y}$ is a vector of the raw phenotypic values for DM, $\mathbf{X}$ is the known incidence matrix for fixed effects β (including grand mean and a nested effect of Trial within Year within Location), $\mathbf{Z}$ is known incidence matrix for clonal additive candidate genomic effects $\mathbf{s}$ and whole genomic effects $\mathbf{g}$. For $\mathbf{K_s}$, and $\mathbf{K_g}$ we used the candidate SNPs and the remaining SNPs from the whole genome excluding the candidate SNPs, respectively, to generate genomic relationship matrices for the 1,651 clones of the GS-C1 population as above. A third kinship matrix, $\mathbf{K_{rand}}$, was generated as a control from 7,203 SNPs anchored to 2000 randomly selected genes from the cassava genome and used in Model (2) in place of $\mathbf{K_s}$, while we calculated $\mathbf{K_g}$ using SNPs from the whole genome excluding those in $\mathbf{K_{rand}}$. Other components of the model include the SnRKs candidate genetic variance σ_s^2 and the genetic variance from other parts of the genome σ_g^2, σ_e^2 is the error variance and e is the residuals from the model. Model (2) was fit using the EMMREML. To assess prediction accuracies, we fit another model as follows:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$$
$$\mathbf{u} \sim N(\mathbf{0}, \sigma_u^2\mathbf{I}); \quad \mathbf{e} \sim N(\mathbf{0}, \sigma_e^2\mathbf{I}) \tag{3.5}$$
$$var(y) = \mathbf{V} = \mathbf{Z}(\sigma_u^2\mathbf{I})\mathbf{Z}^T + \sigma_e^2\mathbf{I}; \quad \hat{\mathbf{u}} = (\sigma_u^2\mathbf{I})\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta})$$

where most components of Model (4.5) remain same as in Model (4.4) apart from the genetic effect **u** having an identity matrix **I** as its covariance matrix signifying that the 1,651 GS-C1 validation clones are unrelated. Model (4.5) was also fit using R EMMREML. Model (4.4) was fit using a 5-fold a cross validation (CV) scheme with 10 repeats and prediction accuracies were obtained for this CV scheme by a correlation of $\hat{\mathbf{s}}$ of each clone from Model (4.4) to its $\hat{\mathbf{u}}$ value from Model (4.5).

Validation using 53 candidate genes extracted from plant physiology literature and 53 randomly selected genes from the RHM significant regions:

We performed a second procedure to validate the 53 candidate genes identified in significant hit regions in the RHM analysis based on plant physiology literature (Table **??**). Using the cassava genome unique gene identifiers from Phytozome (Goodstein et al., 2012), we extracted all markers within 2.5Kb flanking the start and end of each gene as before, resulting in 400 unique SNPs. We refer to these SNPs as likely candidate SNPs. We also picked 53 single copy genes at random from within the RHM significant regions and anchored them to 395 SNPs as controls for the likely candidate SNPs. We term these the unlikely candidate SNPs. To validate these, we also fit the GBLUP Model (2) with these modifications: (1) for $\mathbf{K_s}$ we used $\mathbf{K_{53}}$ which was a genomic relationship matrix calculated from the 400 likely candidate SNPs for the 1,651 clones of the GS-C1 population (as above), (2) we calculated $\mathbf{K_g}$ using SNPs from the whole

genome excluding these likely candidate SNPs, (3) $\mathbf{K_{rand}}$ was also calculated as above (as a control) from 402 SNPs anchored to 53 randomly selected genes from the cassava genome (with 7.5 kb flanking the start and end of these genes), (4) $\mathbf{K_{unlikely}}$ was calculated from the 395 unlikely candidate SNPs. These were also used in place of $\mathbf{K_s}$ in Model (4.4) with their appropriate $\mathbf{K_g}$ calculated as other SNPs in the genome excluding those in $\mathbf{K_{rand}}$ and $\mathbf{K_{unlikely}}$. Other components of the model were as described for Model (4.4) and prediction accuracies were obtained in the same way. To assess the prediction accuracy of the whole genome SNPs, we also fit a model analogous to Model (4.5) with covariance of u coming from a genomic relationship matrix with whole genome SNPs. We term this the predictive accuracy of the whole genome SNPs.

Validation using all genes within 1Mb of the RHM significant list and an a priori list of starch genes in cassava:

We performed another validation procedure to provide a validation for all the genes identified in the significant hit regions in the RHM analysis, including those shown in Table 3.2 and those not shown because they were not selected on the basis of information from literature. Using the cassava genome unique gene identifiers from Phytozome (Goodstein et al., 2012), we extracted all SNPs within a 1 Mb region centered on each of these candidates using Bedtools resulting in 2,297 SNPs from 650 unique genes. We refer to these SNPs as all RHM region SNPs (RHM-regions). In addition we extracted SNPs anchored to 123 unique genes in the cassava starch pathway compiled by Saithong et al. (2013), resulting in 419 SNPs. We refer to these SNPs as cassava starch SNPs. To validate these SNPs, we fit Model (4.4) using genomic relationship matrices calculated as above from RHM-region and cassava starch SNPs, in place of $\mathbf{K_s}$ with

their appropriate $\mathbf{K_g}$ calculated from remaining SNPs. We also picked 650 single copy genes at random excluding the RHM significant regions and anchored them to approximately 2300 SNPs as controls for the RHM-region and cassava starch SNPs. We refer to these as Random-650 SNPs. We calculated $\mathbf{K_{random-650}}$ using these SNPs and an appropriate $\mathbf{K_g}$. These kernels were also fit in Model (4.4) as $\mathbf{K_s}$ and $\mathbf{K_g}$ respectively. In addition to prediction accuracies from these candidates, we validated genes in the RHM-regions by searching for them in two a priori lists compiled by Saithong et al. (2013) including one for the cassava starch pathway and another for the Calvin Cycle pathway. RHM-region genes that made this list were considered validated.

Assessing the RHM power via the hide-a-causal-SNP procedure:

To validate the RHM procedure, we performed an analysis similar to the classical hide-a-causal-SNP approach as follows:

(a) Chromosomes were divided into 100 SNP segments in sliding windows with 50 SNPs overlapping between adjacent segments.

(b) Five (5) adjacent segments were randomly selected on each chromosome.

(c) On the third segment, effects were added to a random SNP to inflate the phenotypic variance of the DM trait by 10%.

(d) Genomic relationship matrices were made for these segments but for segment 3, the random pseudo-causal SNP was excluded when calculating the genomic relationship matrix.

(e) Subsequently, steps (b) to (d) of the RHM procedure above were carried out, resulting in P-values for these five adjacent segments. Steps (a) to (e) were repeated twelve times, resulting in 216 tests.

(f) We then calculated the P-value from the RHM analysis on our data that corresponded to the LFDR threshold of 0.05 and used this as significance threshold.

(g) The power of the RHM analysis was then calculated as the number of times any of the five segment P-values were significant given the significance threshold from (f) above.

(h) To make a decision on the bounds set for extracting adjacent candidate genes from the M. esculenta genome for a significant segment in the RHM analysis, the number of times either the 1st or 5th segment P-values were significant conditional on the 3rd segment having a higher P-value were also calculated. This reflected how far away adjacent segments captured causal variants.

3.6 Results:

3.6.1 RHM for DM in white and yellow cassava populations:

The genomic heritabilities for DM in white and yellow cassava based on whole genome SNPs were 0.57 and 0.48 respectively. These heritabilities are somewhat higher than those found by Ly et al. (2013), presumably because they worked with more locations and years and thus experienced higher genotype-by-environment interaction. We observed different genetic control patterns for DM in the white and yellow cassava subpopulations as shown by Manhattan plots from the RHM analysis (Figure 3.1). Significant genomic segments for the white cassava DM were observed on chromosomes 1, 4, 5, 10, 17 and 18 while

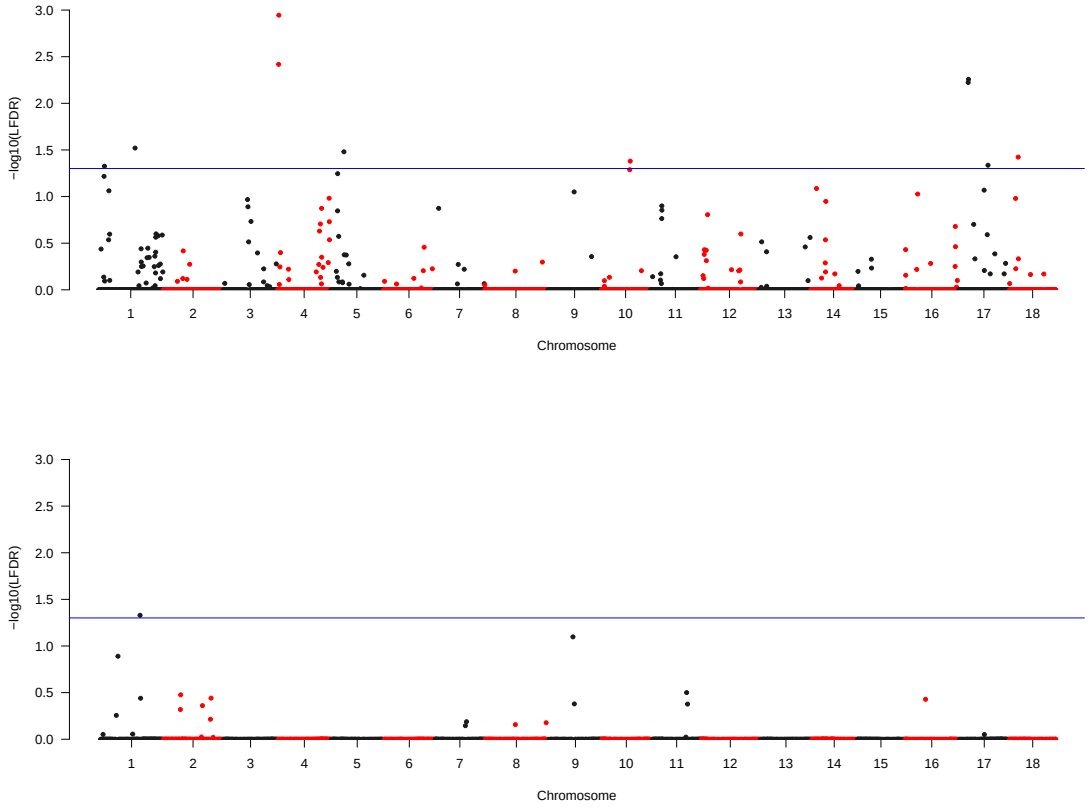

Figure 3.1: **Manhattan plots showing dry matter content genomic segment associations.**

Upper and lower figures show RHM discovery associations for white and yellow cassava populations, respectively.

for the yellow cassava a significant segment was only observed on chromosome 1 (Figure 3.1). Due to the difference between the sample sizes of both subpopulations, it is unclear if the DM genetic control patterns between these subpopulations were different. A non-significant but strong signal was also observed on chromosome 9 of both cassava subpopulations.

3.6.2 Candidate gene analysis:

Using information from the estimates of the mean LD between genomic segments per chromosome (Figure 3.2), the distribution of the length of genomic segments in our analysis (Figure 3.3) and information on the number of times adjacent segments captured causal variations in the simulation analysis; we set the bound for the region where candidate genes were sought to 1.0 Mb (500Kb flanking each hit), representing from two to three genomic segments adjacent to the top hit genomic segment.

3.6.3 Candidates for the white and yellow cassava subpopulations:

For the top RHM hits in both cassava gene pools, we identified possible candidate genes and transcriptional regulators adjacent to these hits based on their involvement in the carbohydrate biosynthesis pathway including members of the serine/threonine kinase family (SnRKs), members of the UDP-glycosyltransferase family (including starch and sucrose synthases), and UDP-sugar transporters, specific plant transcriptional factors including members of the beta helix-loop-helix (bHLH) family and mini zinc fingers, and other genes involved in cell wall processes, root storage and development including pectinases and beta vacuolar processing enzymes. We show a list of these genes in Table 3.2. An additional candidate gene, phosphofructokinase, was associated with the non-significant peak on chromosome 9 which was more pronounced in the yellow cassava germplasm.

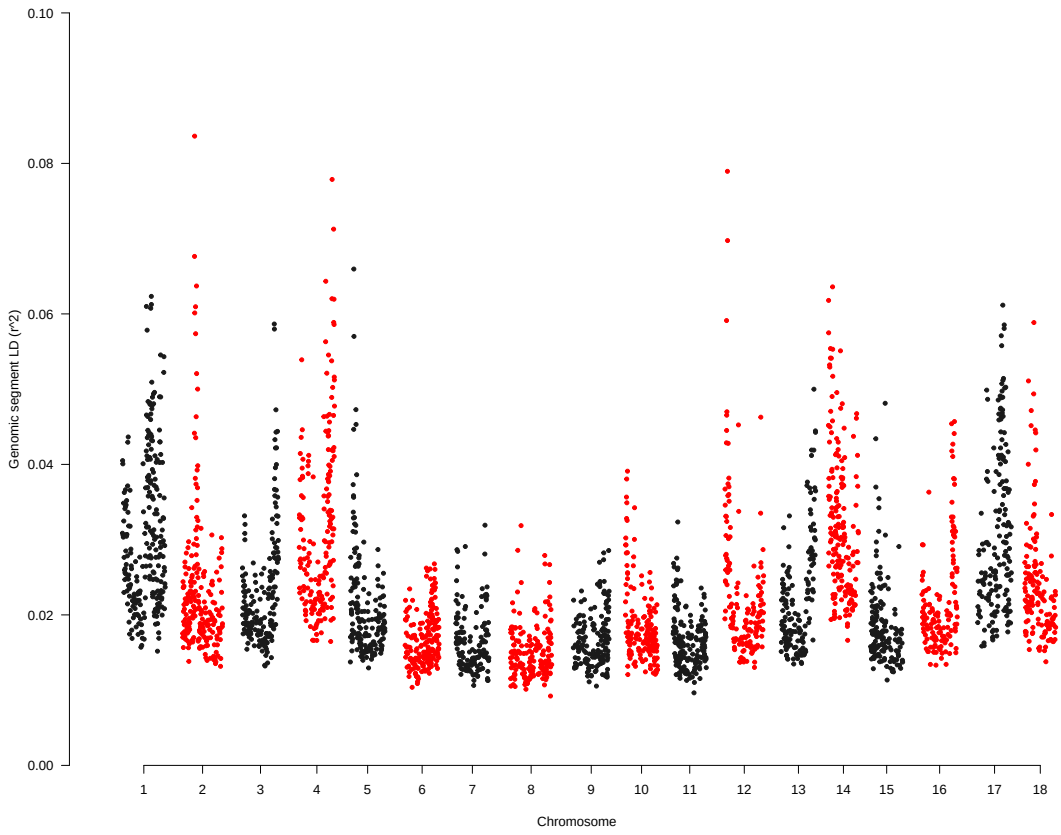

Figure 3.2: **Genome-wide linkage disequilibrium between segments in the RHM analysis.**

Linkage disequilibrium is measured as the mean correlation between all pairs of SNPs where one SNP is on one segment and the other is on the adjacent segment.

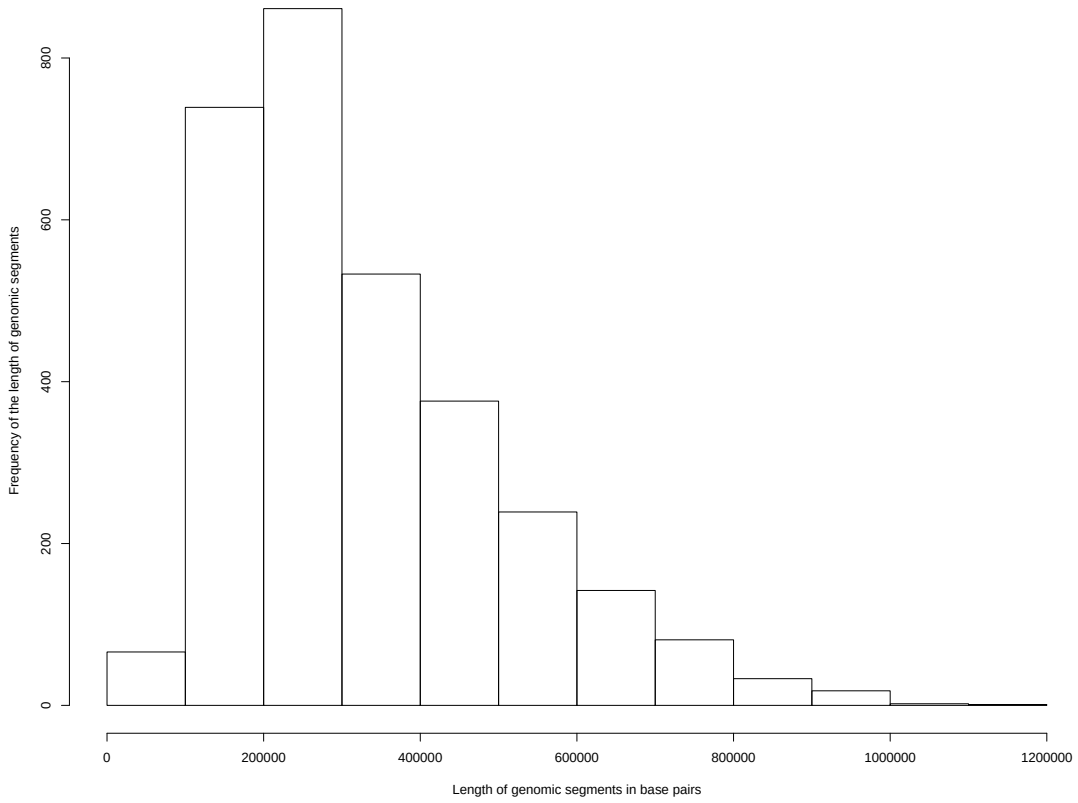

Figure 3.3: **Histogram of the size of genomic segments in the RHM analysis.**

The size of the window is the physical distance in base pairs between the first and the last of the 100 SNPs in the window.

3.6.4 Validation results for SnRKs:

The predictive accuracy of the whole genome SNPs was 0.54 (0.03). Using the set of candidate SnRK SNPs, prediction accuracies from the CV using Model (2) were 0.26 (0.04) and 0.12 (0.06) for the candidate and random SNPs, respectively, with standard deviation of the cross validation repeat cycles shown in parentheses. The predictive ability of the genome-wide SnRK candidates (7,203

SNPs) had approximately 50 percent of the total prediction accuracy from our set of genome-wide SNPs (177,201) for the GS-C1 population.

3.6.5 Validation using 53 likely candidate genes extracted from plant physiology literature and 53 unlikely candidate genes from the RHM significant regions:

Using the likely candidate SNPs from the genes identified for all the top hit genomic segments genome-wide (shown in Table 3.2), prediction accuracies from the CV using a modified Model (2) were 0.17 (0.03), those for the 53 unlikely genes randomly selected from the top hit genomic segments genome-wide were 0.14 (0.02) and those for the SNPs from random 53 genes from the cassava genome were 0.06 (0.08) with standard deviation of the cross validation repeat cycles in parentheses.

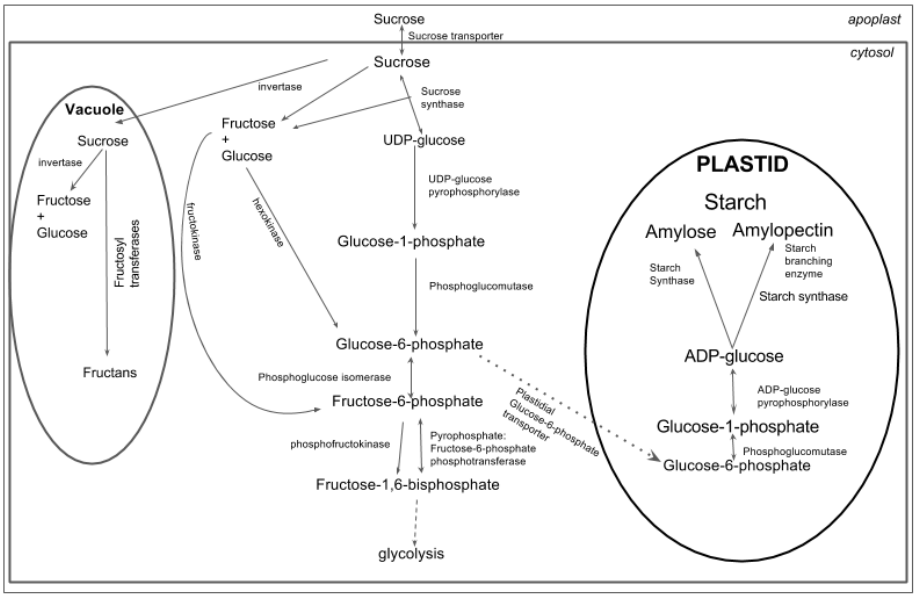

Figure 3.4: **Sucrose/starch metabolism in a heterotrophic plant cell like the cassava tuber.**

Key enzymes including sucrose transporter, invertase, phospho-glucose isomerase, and phosphofructokinase were within 500Kb of significant RHM segments.

3.6.6 Validation using all genes within 1Mb of the RHM significant list and an a priori list of starch genes in cassava:

Using the RHM-region, cassava starch and Random-650 SNPs, the prediction accuracies from the CV using a modified Model (2) were 0.17 (0.04), 0.18 (0.03) and 0.03 (0.01) respectively. Based on two a priori lists compiled by Saithong et al. (2013) including one for the cassava starch pathway and another for the Calvin cycle pathway, we found three RHM-region genes on the cassava starch pathway list including an acid invertase (Manes.01G076500), a glucose-6-phosphate isomerase (Manes.18G060600) and a neutral or alkaline invertase (Manes.04G006900). However, from the Calvin Cycle pathway

list we found one RHM-region gene, namely fructose-biphosphate aldolase (Manes.04G007900). These genes are known to play key roles in starch biosynthesis and storage (Junker, 2004; Ap Rees, 1992; Appeldoorn et al., 1997; Renz et al., 1993). To assess if these genes were significantly enriched in RHM regions, we performed a simple calculation by multiplying the 650 genes in the RHM region with 123 genes in the cassava starch pathway (Saithong et al., 2013) and divided them by the total number of genes in the cassava genome (33,030). The result was 2.4, which is the expectation of a Poisson process of obtaining the genes in the cassava starch pathway. However we calculated the probability of drawing 3 cassava starch pathway genes from the genome at random resulting in p = 0.22 indicating no significant enrichment.

3.6.7 Assessing the RHM power via the hide-a-causal-SNP procedure:

We calculated the statistical power of the RHM procedure to detect simulated causal effects from 216 analyses as the number of times any of the five segment P-values were significant. The P-value from the RHM analysis on our data that corresponded to the LFDR threshold of 0.05 was 0.00024, which became our significance threshold for this analysis. We found that 102 tests were significant out of a total of 216 representing a 47 percent statistical power to detect the simulated causal region. To set the bounds for how far in the genome to cover when extracting candidate genes from an RHM significant segment, we also calculated the number of times P-values from the 1st or 5th genomic segments were significant conditional on the 3rd segments P-value being higher. With a total

Table 3.1: **Summary of validation results for RHM significant candidates.**

Genomic Segment	Prediction accuracy
SnRKs (7,203)	0.26 (0.04)
Random control for SnRKs (7,203)	0.12 (0.06)
53 Likely candidates (400)	0.17 (0.03)
53 Unlikely candidates (395)	0.14 (0.02)
Random control for 53 candidates (402)	0.06 (0.08)
RHM-region genes (2,297)	0.17 (0.04)
Cassava Starch genes (419)	0.18 (0.03)
Random-650 (2,300)	0.03 (0.01)
Whole genome SNPs (177,201)	0.54 (0.03)

Prediction accuracies from selected candidate genes or genomic segments used to validate significance of the RHM hits are given with the number of markers used in each analysis (left column) or the standard deviation of cross validation repeat cycles (right column) in parentheses.

of 216 analysis, 27 cases had significant P-values on segment 3 and 15 cases had significant P-values from segments 1 or 5 when the P-values from segment 3 were higher. This represents 15 percent coverage farther away from the causal segment. With this information we chose an adjacent span of 500,000 kb pairs flanking an RHM significant segment as the bounds for extracting adjacent candidate genes.

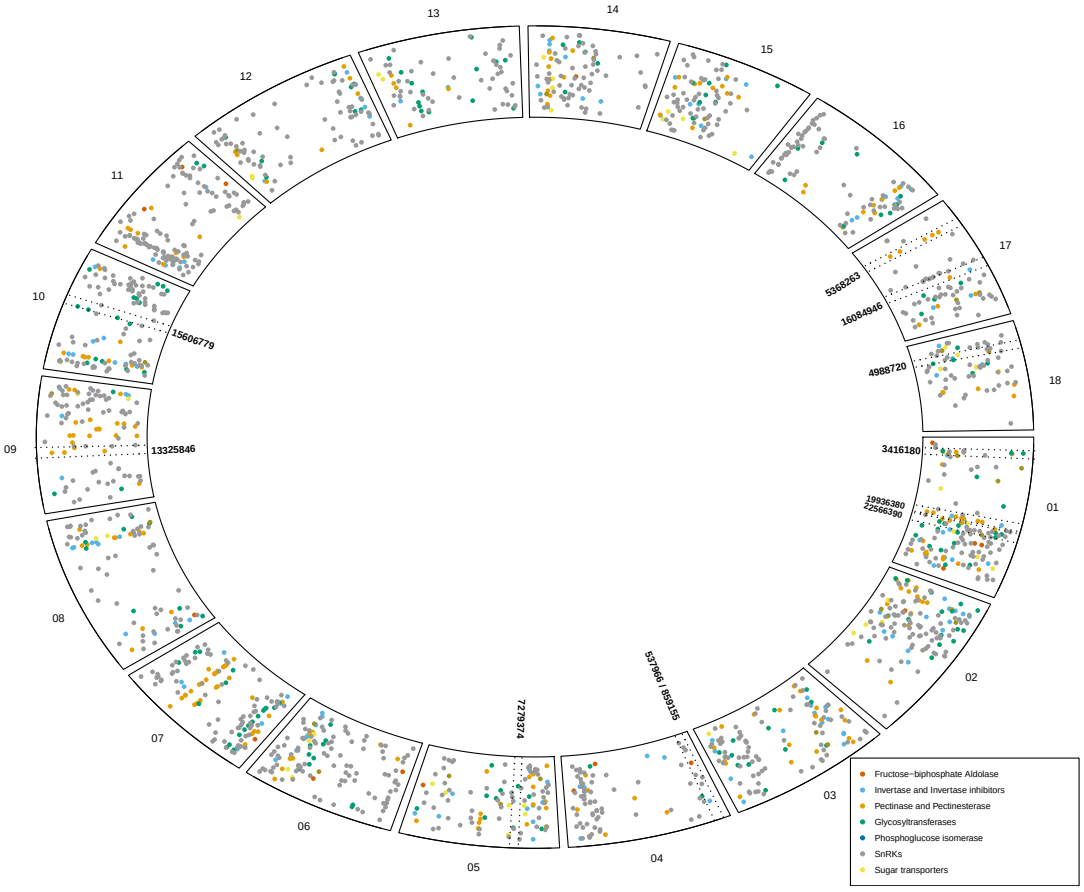

Figure 3.5: **Selected candidate genes and positions of significant RHM segments.**

Circos plot of carbohydrate biosynthesis candidate genes or gene families and significant RHM segments shown by paired dotted lines. Points are randomly scattered along the y-axis to avoid overlaps and better visualize gene families.

3.7 Discussion:

The RHM results in the high DM and white cassava populations clearly demonstrate the polygenic nature of the DM trait. DM is composed of carbohydrates (mostly starch), cell wall components and fiber, as well as other non-starchy polysaccharides. Thus, it is not surprising that this trait is complex and controlled by many genes. Also the RHM procedure in this study showed a 47% power for detection of association with a sample size of less than 500.

3.7.1 SnRKs may be involved in regulation of cassava carbohydrate biosynthesis:

The serine/threonine protein kinase (SnRKs) gene family in plants is homologous to the sucrose non-fermenting 1 (SNF1) protein kinase family in yeast and the AMPK gene family in mammals. Its members have gained recognition as critical elements in transcriptional, metabolic and developmental regulation in plants (Halford et al., 2003; Halford et al., 1998; Polge and Thomas, 2007; Xue-Fei et al., 2012; Crozet et al., 2014; Jossier et al., 2009). The most studied member of this family is the SnRK1 (Halford et al., 1998; Polge and Thomas, 2007). SnRKs play a vital role as global regulators of carbon metabolism and mediate cross talk between metabolic and other plant signaling pathways (Halford et al., 1998; Polge and Thomas, 2007; Xue-Fei et al., 2012). SnRK1 was shown to play a key role in seed filling and maturation and in embryo development in peas (Radchuk et al., 2010; Radchuk et al., 2006). In potato and wheat, SnRK1 phosphorylates and inactivates key enzymes in the sugar and starch biosynthesis

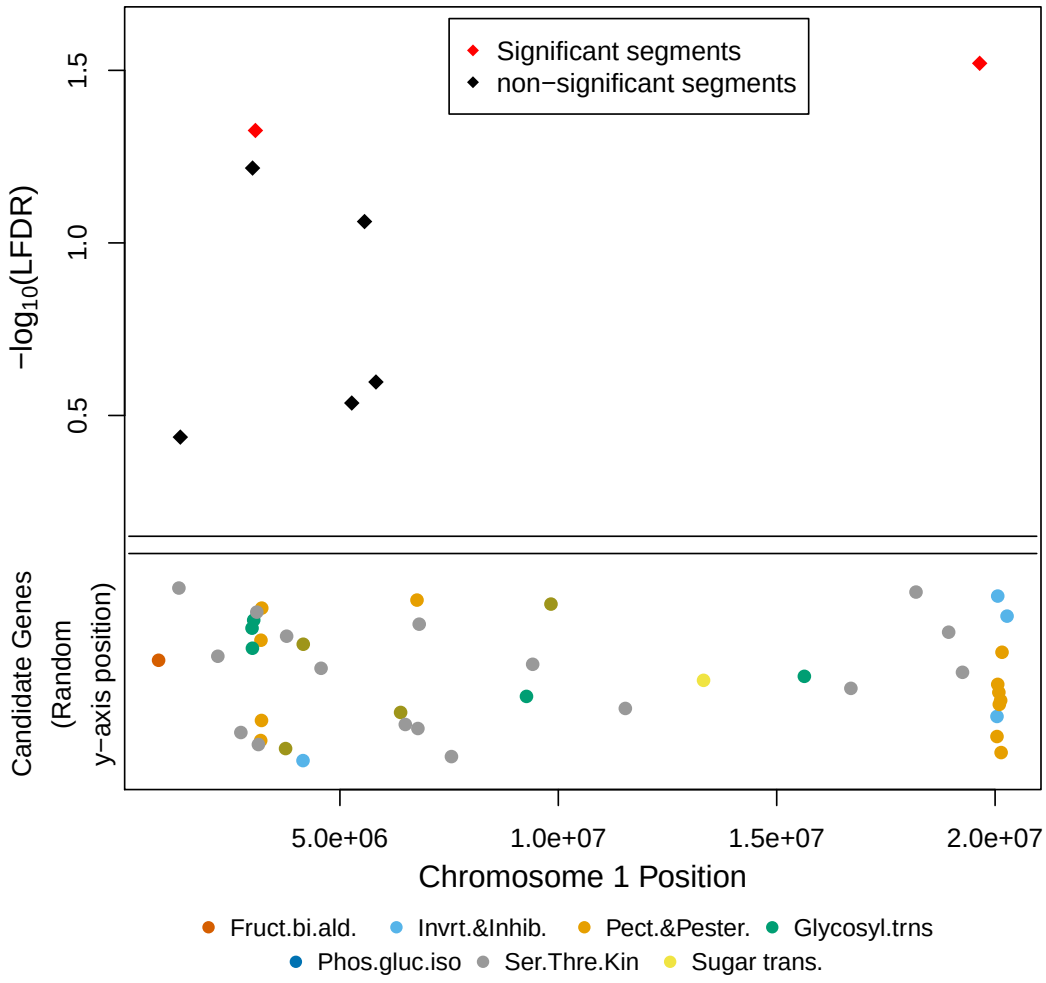

Figure 3.6: **Zoom-in plot of candidate genes and significant RHM segments in a 21Mb region of Chromosome 1.**

The same genes or gene families as Figure 3.5 are shown along with two significant RHM segments. The double line separates candidate genes with random y-axis positions from log 10 (LFDR) plotting of the significance of RHM segments.

pathway, affecting sucrose synthase, trehalose phosphate synthase and alpha amylase (Purcell et al., 1998; Laurie et al., 2003), and in potato, it stimulates the redox activation of ADP-glucose pyrophosphorylase (AGPase) in response to high sucrose levels (Geigenberger, 2003; Tiessen et al, 2003). Antisense expression of SnRK1 resulted in a reduction in the expression of sucrose synthase in potato tubers (Purcell et al., 1998) and alpha amylase in cultured wheat embryos (Laurie et al., 2003). However, the overexpression of SnRK1 in potatoes resulted in a significant increase in starch accumulation in tubers and a decrease in glucose levels resulting from a dramatic increase in the activity and expression levels of sucrose synthase and AGPase (McKibbin et al., 2006). SnRK1 is activated by high cellular sucrose and/ or low glucose or a dark period (Rolland et al., 2002). The model of sugar and starch biosynthesis in potato from McKibbin et al. (2006) showed SnRK1 at the heart of these processes. Using RHM analysis in the white cassava population, we identified significant genomic segments containing some of the proteins or enzymes in the model given in this illustration (McKibbin et al., 2006) including SnRKs, UDP-Glycosyltransferases and UDP-sugar transporters, an ADP-type starch synthase 2 and a neutral invertase. Glycosyltransferases are a family of enzymes involved in carbohydrate biosynthesis of which sucrose and starch synthases are members (Momma and Fujimoto, 2012). Using the RHM procedure and candidate gene analysis, several of these known carbohydrate biosynthesis enzymes (Table 3.2, Figure 3.4) were putatively associated with the cassava DM trait.

3.7.2 Other possible candidates that are involved in sugar and starch biosynthesis in Cassava:

Other proteins located within significant genomic segments that are also involved in the carbohydrate biosynthesis pathway include invertase inhibitors which have been shown to form complexes with SnRKs and lead to reduced accumulation of reducing sugars and increased accumulation of starch in potatoes (Lin et al., 2015), and BAK1, a brassinosteroid insensitive 1 (BR1) associated receptor-like kinase and a member of the somatic embryogenesis receptor-like kinase (SERKs) subfamily involved in regulation of root development (Du et al., 2012). BAK1/serk1 positively controls starch granule accumulation in Arabidopsis root tips (Du et al., 2012). Using a transgenic sweet potato overexpressing a DNA-binding one zinc finger (Dof) protein encoded by a SRF1 gene (a member of the mini zinc finger family of plant specific transcription factors (Takatsuji, 1998; Takatsuji, 1999)), Tanaka et al. (2009) showed that transgenic roots had significantly higher storage root dry matter content, increased starch content per fresh weight of storage root and a drastic decrease in glucose and fructose levels (Tanaka et al., 2009). SRF1 was shown to modulate carbohydrate metabolism in sweet potato storage roots via negative regulation of vacuolar invertase (Tanaka et al., 2009). Several enzymes, including pectinases, pectin esterases, cellulase synthase and galacturonosyltransferases (GAUT), found in the RHM significant regions in white and yellow cassava may be involved in plant cell wall loosening and degradation which may be linked to carbon partitioning in cassava. In fact GAUT, a member of the CAZy (Cantarel et al., 2009) GT8 family of glycosyltransferases, is involved in pectin and hemicellulose biosynthesis (Cantarel et al., 2009; Atmodjo et al., 2011; de Godoy et al., 2013). GAUT-

silenced tomato fruits showed altered pectin composition and decreased starch accumulation (de Godoy et al., 2013). Cassava GAUTs may interfere with carbon metabolism, partitioning and allocation as seen in tomato (de Godoy et al., 2013). In their expression profile study using samples from different stages of cassava root development, Yang et al. (2011) found a significant up-regulation of these enzymes involved in plant cell wall loosening and degradation. The beta helix-loop-helix (bHLH) family of transcription factors is a large family in plants involved in flavonoid, carotenoid pathway and anthocyanin pigmentation of tuber skin and flesh (from yellow to white and purple) in potato (De Jong et al., 2004; Zhang et al., 2009; Tai et al., 2013) and may interact with sucrose transporter to perform this function (Krgel et al., 2012). Phytochrome-interacting factors (PIFs) form a subfamily of bHLH transcription factors and PIF1 (a member of this subfamily) have been shown to directly regulate the expression of phytoene synthase (PSY) (Toledo-Ortiz et al., 2010), a major driver of carotenoid production in plants and the first and main rate-determining enzyme of the carotenoid pathway (Toledo-Ortiz et al., 2010; Maass et al., 2009). It is not clear how bHLH may link with sugar biosynthesis and transport or play a role in starch accumulation in yellow cassava clones, but this may translate to the frequently observed negative correlation between DM and yellow root flesh color in African cassava (Esuma et al., 2016; Akinwale et al., 2010). Interestingly, cassava breeders in Colombia have not found any negative correlation between carotenoids and DM in their germplasm and in fact have made gains in both traits using a rapid cycling recurrent selection scheme (Ceballos et al., 2013).

3.7.3 Some experimental studies that reflect possible roles of candidate genes in the cassava tuber:

Using the RHM analysis, we identified (Figure 3.4) a number of cassava genes in the heterotrophic plant cell starch/sucrose metabolism pathway (Junker, 2004). We describe a few steps in this pathway, concentrating mostly on where we have identified candidate genes (candidate genes are in braces henceforth with phytozome gene identifiers). After sucrose is imported into the cytosol by a sucrose transporter (Manes.05G099000, Manes.18G054200), it is converted into hexose sugars via two paths involving the enzymes sucrose synthase (shown in the center of Fig. 3.4) and invertase (shown to the left in Figure 3.4) (Manes.04G006900, Manes.01G076500) (Junker, 2004; Ap Rees, 1992; Appeldoorn et al., 1997; Renz et al., 1993). Sucrose transport is much more pronounced in the sink tissues that switch to storage mode (Weschke et al., 2000; Weschke et al., 2003). A transgenic study using sucrose transporter 4-RNAi potato plants showed an increase in tuber yield and starch accumulation, and also induced early tuberization (Chincinska et al., 2008). It is worth noting that the cytosolic neutral invertase tends to play a larger role in sink organs than does the vacuolar acid invertase. Studies on maize null mutants of the cytosolic invertase (Mn1) had miniature seeds due to arrested endosperm development (Miller and Chourey, 1992), while overexpression of Mn1 increased grain yield and starch content (Li et al., 2013). Similar studies in rice, tomato and cotton have also found consistent phenotypes with cytosolic neutral invertase (Wang et al., 2008; Zanor et al., 2009; Wang and Ruan, 2012). Other studies on vacuolar invertase inhibitors showed a significant reduction of cold-induced sweetening in potato tubers (via a reduction in sucrose accumulation in tubers) by restricting the activities of

vacuolar acid invertase (McKenzie et al., 2013; Brummell et al., 2011). These studies suggest the importance of sucrose unloading to sink organs and hence vacuolar acid and cytosolic invertases are targets for post-translational regulation towards starch storage and dry matter accumulation (Tang et al., 2016).

The hexoses cleaved from sucrose are rapidly phosphorylated into hexose monophosphates by hexokinase and fructokinase (Junker, 2004; Ap Rees, 1992; Appeldoorn et al., 1997; Renz et al., 1993) and they proceed to starch biosynthesis or glycolytic pathways. As shown in the central pathway in Figure 3.4, the resulting hexose monophosphates (including glucose-1-phosphate, glucose-6-phosphate and fructose-6-phosphate) are interconverted by the enzymes phosphoglucose mutase and phosphoglucose isomerase (Manes.18G060600) (Junker, 2004). Phosphoglucose isomerase connects the Calvin Cycle pathway with the starch biosynthetic pathway in illuminated plant leaves (Bahaji et al., 2015). It also plays a key role in the glycolytic pathway and in the regeneration of glucose-6-phosphate in the oxidative pentose pathway in heterotrophic organs and non-illuminated plant leaves (Bahaji et al., 2015). It is strongly inhibited by light (Heuer et al., 1982) and by an intermediate Calvin Cycle molecule 3-phosphoglycerate (3PGA) (Dietz, 1985), which accumulates in the chloroplast during illumination and allosterically activates AGPase (Kleczkowski, 1999; Kleczkowski, 2000). The second phosphorylation step in the glycolytic pathway is the phosphorylation of fructose-6-phosphate to fructose-1,6-bisphosphate by phosphofructokinase (Manes.09G077800). Interestingly, transgenic studies overexpressing 6-phosphofructokinase in potato found no changes in the transgenic tuber phenotype compared to the controls but had an increased flux of cytosolic 3PGA that did not affect the amount of starch that accumulated in the tubers (Sweetlove et al., 2001; Burrell et al., 1994). It is noteworthy that our RHM

results identified a signal on chromosome 9 in both yellow and white cassava that corresponds to the position of a phosphofructokinase in cassava.

Fructose-bisphosphate aldolase (FDA), a candidate from the Calvin Cycle pathway (Manes.04G007900), is known to play a key role in carbohydrate biosynthesis. Changes in FDA activity have marked consequences for photosynthesis, carbon partitioning, growth, yield and improved uniformity of solids in potato and other plants (Haake et al., 1998; Barry et al., 2002). Transgenic plants (including potato, corn, rice, canola and other crops) that expressed the E. coli FDA gene in their chloroplasts had significantly higher root mass, leaf phenotypes with significantly higher starch accumulation, and lower leaf sucrose compared to control plants expressing the null vector (Barry et al., 2002).

3.7.4 Result implications for the breeding of high DM white cassava varieties or high DM, high beta carotene yellow cassava varieties:

The RHM results presented in this study suggest that DM content is under complex genetic control, particularly in the white cassava population. A network of genes and transcriptional regulons that are at the heart of sugar and starch biosynthesis were positionally associated with significant RHM regions in white and yellow cassava populations. Spurious associations due to linkage may have been avoided in the RHM analysis even when large segments were involved (Figures 3.2 and 3.3). Given the genetic complexity of the cassava DM trait, we suggest that candidate genes, including invertases (neutral and acid) and FDA,

may be targeted for gene editing or transgenic techniques to substantiate the role of these genes in DM and starch accumulation in cassava and to provide a clear path for their utilization in cassava breeding programs.

DM content must work together with fresh root yield (FYLD) to make cassava production profitable and provide value for farmers and processors. To investigate whether some of the genes and gene families identified in the RHM analysis are also involved in the biological processes that lead to cassava FYLD, we validated their effects on FYLD using the same validation procedures and populations as above. The results showed prediction accuracies for SnRKs on FYLD as 0.03 (0.02), 53 likely candidates as 0.02 (0.02), 53 unlikely candidates as 0.006 (0.03), RHM-region genes as 0.03 (0.02), and cassava starch pathway genes as -0.009 (0.02). These results suggest no single biological pathway controls DM and FYLD. This is not surprising since there is little genetic correlation between DM and FYLD (Kawano et al., 1987). It appears from the negative correlation between carotenoid content in roots and DM content in African cassava germplasm (Esuma et al., 2016; Akinwale et al., 2010) and from the link between bHLH and sugar biosynthesis (Krgel et al., 2012), that yellow flesh color is associated with the accumulation of reducing sugars in edible roots (Eleazu and Eleazu, 2012). This poses a more complex challenge for improving DM in African yellow cassava and shifts attention towards finding recombinant yellow cassava progenies that have high DM. Ceballos et al. (2015) states that the search for the appropriate recombinant is difficult in cassava breeding and advocates for the use of inbred progenitors while breeding for hybrid cassava.

In this paper, we have utilized candidate gene analysis attempting to understand the function of the genes or gene families positionally associated with the RHM hits. We do not make the claim that these candidates are causal genes detected by the RHM hits but rather we have shown using prediction accuracies that these RHM hit loci were positionally associated with the DM trait in cassava (Figures 3.1, 3.5 and 3.6) thus resulting in better predictability than random genes used as controls. To validate the hypotheses presented in this paper regarding candidate genes underlying DM accumulation in cassava, and to elucidate the physiological mechanisms involved in the expression of the DM trait in both yellow and white cassava, we recommend the use of genome editing and/or transgenic technology, and in-depth analysis of sugars and carbohydrates in cassava roots, stems and leaves. Similar studies in potato have benefited and informed potato breeding, and the same will be true of cassava as new insights become available.

3.7.5 Conclusion:

Using RHM analysis, we demonstrate the complex genetic architecture of DM content in high DM white African cassava. Candidate gene analysis revealed possible roles of SnRKs, vacuolar and neutral invertases, phosphoglucose isomerase and FDA in the regulation of sugar and starch biosynthesis in cassava. The RHM analysis indicated that inheritance of DM content in the white cassava is polygenic. We examined the utility of models based on genome-wide candidate genes found in this study using prediction accuracies in a different but related population and found appreciable predictive ability compared to what is obtained when whole genome markers were used. Transcriptional regulators

such as bHLH may be involved in flesh root color and sugar biosynthesis in cassava, as shown in potato. We recommend further studies using genome editing or transgenic technology to better understand these mechanisms and to inform and accelerate breeding efforts for cassava.

Table 3.2: **Candidate genes and gene families associated with significant RHM regions**

White cassava DM											
Chromosome	**Segment tag SNP_ID**	**Segment tag SNP bp**	**Segmant Span**	**LFDR**	**Target Segment**	**Genomic**	**Candidate Genes**	**Phytozome ID**	**start (bp)**	**end (bp)**	**homologs**
1	S1_3416180	3416180	3064700 - 3416180	0.047227533	0.66	0.84	BAK1-interacting receptor-like kinase 1 (BIR1)	Manes.01G018900	3095915	3098209	AT5G48380.1
							BAK1-interacting receptor-like kinase 1 (BIR1)	Manes.01G019000	3132412	3134588	AT5G48380.1
1	S1_19936380	19936380	19646964 - 19936380	0.030168971	0.58	0.82	UDP-D-glucuronate 4-epimerase 2	Manes.01G074000	19949560	19951824	AT1G02000.1
							DHHC-type zinc finger family protein	Manes.01G073100	19818085	19825785	AT3G60800.1
							Bifunctional inhibitor trypsin-alpha amylase inhibitor	Manes.01G074600	20036878	20037858	AT1G62790.2
							Plant invertase/pectin methylesterase inhibitor superfamily protein	Manes.01G074800	20042476	20043982	AT4G25260.1
							Pectinesterase	Manes.01G075600	20158111	20159120	AT5G07420.1
							beta vacuolar processing enzyme	Manes.01G075700	20173648	20176768	AT1G62710.1
							mini zinc finger 2	Manes.01G090800	21547715	21548005	AT3G28917.1
							galactosyltransferase1	Manes.01G091600	21599298	21603804	AT1G26810.1
							starch synthase 2 (ADP-Glucose type)	Manes.01G091700	21623316	21629007	AT3G01180.1
							cycling DOF factor 2	Manes.01G092100	21647142	21650438	AT5G39660.1
4	S4_537966	537966	510971 - 732069	0.003820074	0.67	0.84	serine/threonine-protein kinase RIO	Manes.04G004800	535401	538252	AT1G08290.1
							Glycosyltransferase family 29	Manes.04G004900	542055	543236	AT1G08280.1
4	S4_859155	859155	623461 - 859155	0.00113334	0.62	0.82	alkaline/neutral invertase	Manes.04G006900	778931	784740	AT5G22510.1
							cellulose synthase-like A02	Manes.04G009400	1064431	1069211	AT5G22740.1
							C2H2-like zinc finger protein	Manes.04G008800	962985	963932	AT4G35280.1
							serine/threonine protein kinase (SnRK1)	Manes.04G006600	757312	760778	AT3G44610.1
5	S5_7279374	7279374	7061162 - 7279374	0.03310678	0.69	0.85	trehalose-phosphatase/synthase 9	Manes.05G087900	6905372	6911159	AT1G23870.1
							Serine/threonine-protein kinase WNK (With No Lysine)-related	Manes.05G089000	7284514	7286444	AT1G60060.1
							serine/threonine protein kinase (SnRK1)	Manes.05G090100	7365593	7370118	AT1G24030.2
							NAD(P)-linked oxidoreductase superfamily protein	Manes.05G092400	7630814	7632642	AT1G59960.1
							galactosyltransferase family protein	Manes.05G095000	7849628	7853700	AT5G62620.1
							bHLH Transcription factor	Manes.05G094700	7824312	7824497	LOC_Os07g09590.1
							Galactose oxidase	Manes.05G096300	7992476	7993801	Cre06.g306000.t1.1
							alpha-amylase-like 3	Manes.05G097100	8094930	8103684	AT1G69830.1
							sucrose transport protein (SUC3)	Manes.05G099000	8344933	8345349	LOC_Os02g36700.1
							UDP-galactose/UDP-glucose transporter	Manes.05G101600	8633269	8636565	Potri.016G139100.1
							zinc finger, C3HC4 type domain containing protein	Manes.05G102000	8696530	8696856	LOC_Os03g20870.1
10	S10_15606779	15606779	15024128 - 15606779	0.041702151	0.15	0.64	serine/threonine protein kinase (SnRK1)	Manes.10G079500	14844954	14851620	AT1G49730.1
							basic helix-loop-helix (bHLH) DNA-binding superfamily protein	Manes.10G080300	15279304	15281955	AT1G49770.1
							salt tolerance zinc finger	Manes.10G080000	15071612	15072680	AT1G27730.1
							Protein with RNI-like/FBD-like domains	Manes.10G080200	15240978	15241397	AT5G56410.1
							Regulator of Vps4 activity in the MVB pathway protein	Manes.10G080400	15346375	15347741	AT2G14830.2
17	S17_5368263	5368263	5152794 - 5368263	0.005975712	0.17	0.64	xyloglucan endotransglucosylase/hydrolase 30	Manes.17G015100	5198786	5201174	AT1G32170.1
17	S17_5532660		5285956 - 5532660	0.005529599	0.27	0.67	Pectin lyase-like superfamily protein	Manes.17G015200	5264305	5276255	AT3G07840.1
							Basic helix-loop-helix (bHLH) DNA-binding family protein	Manes.17G016000	5657214	5660207	AT1G32640.1
17	S17_16084946	16084946	15509115 - 16084946	0.046144077	0.14	0.62	CBL-interacting protein kinase 23 (SnRK3)	Manes.17G073900	21282260	21287846	AT1G30270.1
							inositol transporter 2	Manes.17G073000	21211865	21215459	AT1G30220.1
							Transducin/WD40 repeat-like superfamily protein	Manes.17G073300	21251484	21254389	AT1G65030.1
18	S18_4988720	4988720	4586953 - 4988720	0.037726745	0.69	0.85	sucrose transporter 4	Manes.18G054200	4548075	4559586	AT1G09960.1
							Nucleotide-diphospho-sugar transferases superfamily protein	Manes.18G054400	4572586	4576353	AT1G27600.1
							beta HLH protein 93	Manes.18G055000	4655203	4657015	AT5G65640.1
							CBL-interacting protein kinase 8 (SnRK3)	Manes.18G055300	4677895	4683026	AT4G24400.1
							UDP-Glycosyltransferase superfamily protein	Manes.18G056200	4793379	4795015	AT5G49690.1

Yellow DM											
Chromosome	**Segment tag SNP_ID**	**Segment tag SNP bp**	**Segmant Span**	**LFDR**	**Target Segment**	**Genomic**	**Candidate Genes**	**Phytozome ID**	**start (bp)**	**end (bp)**	**homologs**
1	S1_22566390	22566390	21910139 - 22566390	0.04682523	0.72	0.87	galacturonosyltransferase 13	Manes.01G098400	22245408	22256933	AT3G01040.1
							Plant calmodulin-binding protein-related	Manes.01G099000	22282570	22286073	AT5G39380.1
							UDP-glucosyl transferase 76E11	Manes.01G100300	22370079	22371643	AT3G46670.1
							SNF1 kinase (SnRK1)	Manes.01G100900	22409977	22417397	AT3G01090.2
							Vesicle transport v-SNARE family protein	Manes.01G101200	22421446	22425558	AT5G39510.1
							brassinosteroid-6-oxidase 2	Manes.01G102800	22514594	22518856	AT3G30180.1
*9	S9_13325846	13325846	12810619 - 13747638		0.26	0.66	phosphofructokinase 2 (PFK2)	Manes.09G077800	13223145	13226580	AT5G47810.1

* Strong but non-significant signal

3.8 Declarations:

Authors contributions:

UGO designed, carried out study and drafted manuscript, DA provided statistical assistance and advice, JLJ supervised the study, designed the validation procedures and revised manuscript, IR and PK supervised data generation for this study and revised manuscript. All authors read and approved manuscript.

Acknowledgements:

We acknowledge the Bill and Melinda Gates Foundation and UKaid (Grant 1048542; http://www.gatesfoundation.org) and support from the CGIAR Research Program on Roots, Tubers and Bananas (http://www.rtb.cgiar.org). Deniz Akdemir was supported by the USDA-NIFA-AFRI Triticeae Coordinated Agricultural Project, award number 2011-68002-30029. We give special thanks to Professor Susan McCouch for her technical review and insights. Thanks also to A. I. Smith and technical teams at IITA for collection of other phenotypic data and to A. Agbona and M. Wolfe for data curation.

Competing interests:

The authors declare that they have no competing interests.

3.9 Future directions:

A particularly interesting gene family which was positionally associated with RHM significant regions in this study was SnRKs. SnRKs are regulatory genes well distributed in the cassava genome. Cassava SnRKS might not be causal but just distributed similarly to causal polymorphisms. It would be an interesting study to understand why there are so much SnRKs (1250) in the cassava genome and why they are distributed across all chromosomes. In addition to this, are these SnRKs tagging causal polymorphisms? If we take genes located within 10-50 Kb around each SnRK in the genome, anchor them to SNPs and use them in the RHM for cassava DM, will these have a prediction accuracy close to what was obtained for SnRKs? On the other hand, how will the RHM perform if segments were chosen based on the LD structure of cassava? Can we still detect these same significant regions using LD informed segments? How will these segments be deployed for DM improvement? Can we use the LD-informed RHM procedure to understand quantitative inheritance for cassava CMD and CBSD?

These questions require careful investigations and may lead to better understanding of both the RHM procedure and also cassava DM inheritance.

3.10 References:

Adewusi, S.R. and Bradbury, J.H., 1993. Carotenoids in cassava: Comparison of opencolumn and HPLC methods of analysis. Journal of the Science of Food and Agriculture, 62(4), pp.375-383.

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

Aniedu, C. and Omodamiro, R.M., 2012. Use of Newly Bred -Carotene Cassava in Production of Value-Added Products: Implication for Food Security in Nigeria. Global Journal of Science Frontier Research, 12(10-D).

Ap Rees, T., 1992. Synthesis of storage starch. Carbon partitioning within and between organisms. Bios Scientific Publishers, Oxford, pp.115-131.

Appeldoorn, N.J., de Bruijn, S.M., Koot-Gronsveld, E.A., Visser, R.G., Vreugdenhil, D. and van der Plas, L.H., 1997. Developmental changes of enzymes involved in conversion of sucrose to hexose-phosphate during early tuberisation of potato. Planta, 202(2), pp.220-226.

Atmodjo, M.A., Sakuragi, Y., Zhu, X., Burrell, A.J., Mohanty, S.S., Atwood, J.A., Orlando, R., Scheller, H.V. and Mohnen, D., 2011. Galacturonosyltransferase (GAUT) 1 and GAUT7 are the core of a plant cell wall pectin biosynthetic homogalacturonan: galacturonosyltransferase complex. Proceedings of the National Academy of Sciences, 108(50), pp.20225-20230.

Ayres, D.L., Darling, A., Zwickl, D.J., Beerli, P., Holder, M.T., Lewis, P.O., Huelsenbeck, J.P., Ronquist, F., Swofford, D.L., Cummings, M.P. and Rambaut, A., 2011. BEAGLE: an application programming interface and high-

performance computing library for statistical phylogenetics. Systematic biology, p.syr100.

Bahaji, A., Snchez-Lpez, .M., De Diego, N., Muoz, F.J., Baroja-Fernndez, E., Li, J., Ricarte-Bermejo, A., Baslam, M., Aranjuelo, I., Almagro, G. and Humplk, J.F., 2015. Plastidic phosphoglucose isomerase is an important determinant of starch accumulation in mesophyll cells, growth, photosynthetic capacity, and biosynthesis of plastidic cytokinins in Arabidopsis. PloS one, 10(3), p.e0119641.

Barrios, E.A., and R. Bressani. 1967. Composicion quimica de la raiz y de la hoja de algunas variedades de yuca Manihot. Turrialba 17:314-320.

Barry, G.F., Cheikh, N. and Kishore, G.M., Monsanto Technology Llc, 2002. Expression of fructose 1, 6 bisphosphate aldolase in transgenic plants. U.S. Patent 6,441,277.

Bland, J.M. and Altman, D.G., 1995. Multiple significance tests: the Bonferroni method. Bmj, 310(6973), p.170.

Bodmer, W. and Tomlinson, I., 2010. Rare genetic variants and the risk of cancer. Current opinion in genetics and development, 20(3), pp.262-267.

Bouis, H.E., Hotz, C., McClafferty, B., Meenakshi, J.V. and Pfeiffer, W.H., 2011. Biofortification: a new tool to reduce micronutrient malnutrition. Food and nutrition bulletin, 32(1 suppl1), pp.S31-S40.

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. and Nauluvula, P., 2016. Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity.Nature biotechnology, 34(5), pp.562-570.

Brummell, D.A., Chen, R.K., Harris, J.C., Zhang, H., Hamiaux, C., Kralicek, A.V. and McKenzie, M.J., 2011. Induction of vacuolar invertase inhibitor mRNA in potato tubers contributes to cold-induced sweetening resistance and includes spliced hybrid mRNA variants. Journal of experimental botany, 62(10), pp.3519-3534.

Burrell, M.M., Mooney, P.J., Blundy, M., Carter, D., Wilson, F., Green, J., Blundy, K.S. and Rees, T.A., 1994. Genetic manipulation of 6-phosphofructokinase in potato tubers. Planta, 194(1), pp.95-101.

Caballero, A., Tenesa, A. and Keightley, P.D., 2015. The Nature of Genetic Variation for Complex Traits Revealed by GWAS and Regional Heritability Mapping Analyses. Genetics, 201(4), pp.1601-1613.

Cantarel, B.L., Coutinho, P.M., Rancurel, C., Bernard, T., Lombard, V. and Henrissat, B., 2009. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. Nucleic acids research, 37(suppl 1), pp.D233-D238.

Ceballos, H., Iglesias, C.A., Prez, J.C. and Dixon, A.G., 2004. Cassava breeding: opportunities and challenges. Plant molecular biology, 56(4), pp.503-516.

Ceballos, H., Kawuki, R.S., Gracen, V.E., Yencho, G.C. and Hershey, C.H., 2015. Conventional breeding, marker-assisted selection, genomic selection and inbreeding in clonally propagated crops: a case study for cassava. Theoretical and Applied Genetics, 128(9), pp.1647-1667.

Ceballos, H., Morante, N., Snchez, T., Ortiz, D., Aragon, I., Chvez, A.L., Pizarro, M., Calle, F. and Dufour, D., 2013. Rapid cycling recurrent selection for increased carotenoids content in cassava roots. Crop Science, 53(6), pp.2342-2351.

Chvez, A.L., Snchez, T., Jaramillo, G., Bedoya, J., Echeverry, J., Bolaos, E.A., Ceballos, H. and Iglesias, C.A., 2005. Variation of quality traits in cassava roots evaluated in landraces and improved clones.Euphytica, 143(1-2), pp.125-133.

Chincinska, I.A., Liesche, J., Krgel, U., Michalska, J., Geigenberger, P., Grimm, B. and Khn, C., 2008. Sucrose transporter StSUT4 from potato affects flowering, tuberization, and shade avoidance response. Plant Physiology, 146(2), pp.515-528.

Crozet, P., Margalha, L., Confraria, A., Rodrigues, A., Martinho, C., Adamo, M., Elias, C.A. and Baena-Gonzlez, E., 2014. Mechanisms of regulation of SNF1/AMPK/SnRK1 protein kinases. Frontiers in plant science, 5, p.190.

de Godoy, F., Bermdez, L., Lira, B.S., de Souza, A.P., Elbl, P., Demarco, D., Alseekh, S., Insani, M., Buckeridge, M., Almeida, J. and Grigioni, G., 2013. Galacturonosyltransferase 4 silencing alters pectin composition and carbon partitioning in tomato. Journal of experimental botany, 64(8), pp.2449-2466.

De Jong, W.S., Eannetta, N.T., De Jong, D.M. and Bodis, M., 2004. Candidate gene analysis of anthocyanin pigmentation loci in the Solanaceae.Theoretical and Applied Genetics, 108(3), pp.423-432.

Deniz Akdemir and Okeke Uche Godfrey (2014). EMMREML: Fitting mixed models with known covariance structures. R package version 2.0. http://CRAN.R- project.org/package=EMMREML

Dietz, K.J., 1985. A possible rate-limiting function of chloroplast hexose-monophosphate isomerase in starch synthesis of leaves. Biochimica et Biophysica Acta (BBA)-General Subjects, 839(3), pp.240-248.

Du, J., Yin, H., Zhang, S., Wei, Z., Zhao, B., Zhang, J., Gou, X., Lin, H. and Li, J., 2012. Somatic embryogenesis receptor kinases control root development mainly via brassinosteroidindependent actions in Arabidopsis thaliana. Journal of integrative plant biology, 54(6), pp.388-399.

Eleazu, C.O. and Eleazu, K.C., 2012. Determination of the proximate composition, total carotenoid, reducing sugars and residual cyanide levels of flours of 6 new yellow and white cassava (Manihot esculenta Crantz) varieties. American Journal of Food Technology, 7(10), pp.642-649.

El-Sharkawy, M.A., 2003. Cassava biology and physiology. Plant molecular biology, 53(5), pp.621-641.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S. and Mitchell, S.E., 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS one, 6(5), p.e19379.

EMBL-EBI InterPro. Protein Sequence analysis and classification. `http://www.ebi.ac.uk/interpro/protein/D1G5D2/similar-proteins;jsessionid=10EFEE974F312EC7435985662EB07A7D`. Accessed October 9, 2016.

Enidiok, S.E., Attah, L.E. and Otuechere, C.A., 2008. Evaluation of moisture, total cyanide and fiber contents of garri produced from cassava (Manihot utilissima) varieties obtained from Awassa in Southern Ethiopia.Pak. J. Nutr, 7, pp.625-629.

Esuma, W., Kawuki, R.S., Herselman, L. and Labuschagne, M.T., 2016. Diallel analysis of provitamin A carotenoid and dry matter content in cassava (Manihot esculenta Crantz). Breeding Science, 66(4), pp.627-635.

Food and Agriculture Organization of the United Nations. FAOSTAT Statistics Database. :FAO, 2013.

Geigenberger, P., 2003. Regulation of sucrose to starch conversion in growing potato tubers. Journal of Experimental Botany, 54(382), pp.457-465.

Gibson, G., 2012. Rare and common variants: twenty arguments. Nature Reviews Genetics, 13(2), pp.135-145.

Glaubitz, J., T. Casstevens, R. Elshire, J. Harriman, and E.S. Buckler. 2012. TASSEL 3.0 genotyping by sequencing (GBS) pipeline documentation. Edward S. Buckler, USDA-ARS, Ithaca, NY. `http://www.maizegenetics.net/tassel/docs/TasselPipelineGBS.pdf` (accessed 3. Jan. 2014)

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., ... and Rokhsar, D. S. (2012). Phytozome: a comparative platform for green plant genomics. Nucleic acids research, 40(D1), D1178-D1186. (http://www.phytozome.net; accessed 1 July, 2014).

Haake, V., Zrenner, R., Sonnewald, U. and Stitt, M., 1998. A moderate decrease of plastid aldolase activity inhibits photosynthesis, alters the levels of sugars and starch, and inhibits growth of potato plants. The Plant Journal, 14(2), pp.147-157.

Halford, N.G. and Hardie, D.G., 1998. SNF1-related protein kinases: global regulators of carbon metabolism in plants?. Plant molecular biology, 37(5), pp.735-748.

Halford, N.G., Hey, S., Jhurreea, D., Laurie, S., McKibbin, R.S., Paul, M. and Zhang, Y., 2003. Metabolic signalling and carbon partitioning: role of Snf1related (SnRK1) protein kinase. Journal of Experimental Botany, 54(382), pp.467-475.

Heuer, B., Hansen, M.J. and Anderson, L.E., 1982. Light modulation of phosphofructokinase in pea leaf chloroplasts. Plant physiology, 69(6), pp.1404-1406.

Holden, H.M., Rayment, I. and Thoden, J.B., 2003. Structure and function of enzymes of the Leloir pathway for galactose metabolism. Journal of Biological Chemistry, 278(45), pp.43885-43888.

Holleman, L.H. and Aten, A., 1956.Processing of cassava and cassava products in rural industries(No. 04; HD9235. C36, H6.). P. imprenta: Rome, FAO, 1956, 115 p. ill. (FAO Agricultural developed paper, 54). http://www.sidalc.net/cgi-bin/wxis.exe/?IsisScript=COLPOS.xis&method=post&formato=2&cantidad=1&expresion=mfn=018400

Iglesias, Carlos .A. and Hershey, Clair .H. (1994). TRUE CASSAVA SEED: RESEARCH FOR A PRODUCTION ALTERNATIVE. Acta Hortic. 380, 164-171. DOI: 10.17660/ActaHortic.1994.380.24. https://doi.org/10.17660/ActaHortic.1994.380.24

Iglesias, C., Hershey, C., Calle, F. and Bolaos, A., 1994. Propagating cassava (Manihot esculenta) by sexual seed.Experimental Agriculture,30(3), pp.283-290.

John D. Storey with contributions from Andrew J. Bass, Alan Dabney and David Robinson (2015). qvalue: Q-value estimation for false discovery rate control. R package version 2.2.2. http://github.com/jdstorey/qvalue (Accessed May 5, 2016).

Jossier, M., Bouly, J.P., Meimoun, P., Arjmand, A., Lessard, P., Hawley, S., Grahame Hardie, D. and Thomas, M., 2009. SnRK1 (SNF1related kinase 1)

has a central role in sugar and ABA signalling in Arabidopsis thaliana. The Plant Journal, 59(2), pp.316-328.

Junker, B.H., 2004. Sucrose breakdown in the potato tuber. Mathematisch-Naturwissenschaftliche Fakultt. Universitt Potsdam, 126.

Kawano, K., Fukuda, W.M.G. and Cenpukdee, U., 1987. Genetic and environmental effects on dry matter content of cassava root. Crop Science, 27(1), pp.69-74.

Kawano, K., Fukuda, W.M.G. and Cenpukdee, U., 1987. Genetic and environmental effects on dry matter content of cassava root. Crop Science,27(1), pp.69-74.

Kawano, K., Narintaraporn, K., Narintaraporn, P., Sarakarn, S., Limsila, A., Limsila, J., Suparhan, D., Sarawat, V. and Watananonta, W., 1998. Yield improvement in a multistage breeding program for cassava. Crop Science, 38(2), pp.325-332.

Keating, B.A., Wilson, G.L. and Evenson, J.P., 1988. Effects of length, thickness, orientation, and planting density of cassava (Manihot esculenta Crantz) planting material on subsequent establishment, growth and yield. E. Afr. Agric. For. J, 53, pp.145-149.

Kimura, M., Kobori, C.N., Rodriguez-Amaya, D.B. and Nestel, P., 2007. Screening and HPLC methods for carotenoids in sweetpotato, cassava and maize for plant breeding trials. Food Chemistry, 100(4), pp.1734-1746.

Kleczkowski, L.A., 1999. A phosphoglycerate to inorganic phosphate ratio is the major factor in controlling starch levels in chloroplasts via ADPglucose pyrophosphorylase regulation. FEBS letters, 448(1), pp.153-156.

Kleczkowski, L.A., 2000. Is leaf ADP-glucose pyrophosphorylase an allosteric enzyme?. Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology, 1476(1), pp.103-108.

Klein, R.J., 2007. Power analysis for genome-wide association studies. BMC genetics, 8(1), p.58. Krgel, U., He, H.X., Gier, K., Reins, J., Chincinska, I., Grimm, B., Schulze, W.X. and Khn, C., 2012. The potato sucrose transporter StSUT1 interacts with a DRM-associated protein disulfide isomerase. Molecular plant, 5(1), pp.43-62.

La Frano, M.R., Woodhouse, L.R., Burnett, D.J. and Burri, B.J., 2013. Biofortified cassava increases -carotene and vitamin A concentrations in the TAG-rich plasma layer of American women. British Journal of Nutrition,110(02), pp.310-320.

Laurie, S., McKibbin, R.S. and Halford, N.G., 2003. Antisense SNF1related (SnRK1) protein kinase gene represses transient activity of an amylase (Amy2) gene promoter in cultured wheat embryos. Journal of Experimental Botany, 54(383), pp.739-747.

Li, B., Liu, H., Zhang, Y., Kang, T., Zhang, L., Tong, J., Xiao, L. and Zhang, H., 2013. Constitutive expression of cell wall invertase genes increases grain yield and starch content in maize. Plant biotechnology journal, 11(9), pp.1080-1091.

Lim, H.K. 1968. Composition data of feeds and concentrates. Malay.Agric.J.46:63-79.

Lin, Y., Liu, T., Liu, J., Liu, X., Ou, Y., Zhang, H., Li, M., Sonnewald, U., Song, B. and Xie, C., 2015. Subtle regulation of potato acid invertase activity by a protein complex of invertase, invertase inhibitor, and SU-

CROSE NONFERMENTING1-RELATED PROTEIN KINASE. Plant physiology,168(4), pp.1807-1819.

Liu, W., Zhou, Y., Sanchez, T., Ceballos, H. and White, W.S., 2010. The vitamin A equivalence of beta-carotene in beta-carotene-biofortified cassava ingested by women. The FASEB Journal, 24(1_MeetingAbstracts), pp.92-7.

Long, A.D. and Langley, C.H., 1999. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Research, 9(8), pp.720-731.

Ly, Delphine, et al.; Relatedness and genotype environment interaction affect prediction accuracies in genomic selection: A study in cassava. Crop Science53.4 (2013): 1312-1325.

Maass, D., Arango, J., Wst, F., Beyer, P. and Welsch, R., 2009. Carotenoid crystal formation in Arabidopsis and carrot roots caused by increased phytoene synthase protein levels. PLoS One, 4(7), p.e6373.

Maziya-Dixon, B., A.G.O. Dixon, and A.-R.A. Adebowale. 2007. Targeting different end uses of cassava: Genotypic variations for cyanogenic potentials and pasting properties. Int. J. Food Sci. Technol. 42:969976. doi:10.1111/j.1365-2621.2006.01319.x

Mckenzie, M.J., Chen, R.K., Harris, J.C., Ashworth, M.J. and Brummell, D.A., 2013. Posttranslational regulation of acid invertase activity by vacuolar invertase inhibitor affects resistance to coldinduced sweetening of potato tubers. Plant, cell and environment, 36(1), pp.176-185.

McKibbin, R.S., Muttucumaru, N., Paul, M.J., Powers, S.J., Burrell, M.M., Coates, S., Purcell, P.C., Tiessen, A., Geigenberger, P. and Halford, N.G., 2006. Production of highstarch, lowglucose potatoes through overexpres-

sion of the metabolic regulator SnRK1. Plant biotechnology journal,4(4), pp.409-418.

Meenakshi, J.V., Banerji, A., Manyong, V., Tomlins, K., Hamukwala, P., Zulu, R. and Mungoma, C., 2010. Consumer acceptance of provitamin A orange maize in rural Zambia (HarvestPlus Working Paper No. 4).

Meenakshi, J.V., Johnson, N.L., Manyong, V.M., DeGroote, H., Javelosa, J., Yanggen, D.R., Naher, F., Gonzalez, C., Garca, J. and Meng, E., 2010. How cost-effective is biofortification in combating micronutrient malnutrition? An ex ante assessment. World Development, 38(1), pp.64-75.

Miller, M.E. and Chourey, P.S., 1992. The maize invertase-deficient miniature-1 seed mutation is associated with aberrant pedicel and endosperm development. The Plant Cell, 4(3), pp.297-305.

Mitchell, A., Chang, H.Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., McAnulla, C., McMenamin, C., Nuka, G., Pesseat, S. and Sangrador-Vegas, A., 2014. The InterPro protein families database: the classification resource after 15 years. Nucleic acids research, p.gku1243.

Momma, M. and Fujimoto, Z., 2012. Interdomain disulfide bridge in the rice granule bound starch synthase I catalytic domain as elucidated by X-ray structure analysis. Bioscience, biotechnology, and biochemistry, 76(8), pp.1591-1595.

Nagamine, Y., Pong-Wong, R., Navarro, P., Vitart, V., Hayward, C., Rudan, I., Campbell, H., Wilson, J., Wild, S., Hicks, A.A. and Pramstaller, P.P., 2012. Localising loci underlying complex trait variation using regional genomic relationship mapping. PloS one, 7(10), p.e46501.

Okechukwu, R.U., and A.G.O. Dixon. 2008. Genetic gains from 30 years of cassava breeding in Nigeria for storage root yield and dis-

ease resistance in elite cassava genotypes. J. Crop Improv. 22:181208. doi:10.1080/15427520802212506

Pfeiffer, W.H. and McClafferty, B., 2007. HarvestPlus: breeding crops for better nutrition. Crop Science, 47(Supplement_3), pp.S-88.

Plus, H., 2009. Breeding crops for better nutrition. Provitamin A Cassava. http://r4d.dfid.gov.uk/PDF/Outputs/Misc_Crop/HarvstPlus_Cassava_Strategy.pdf. Accessed June 30, 2016.

Polge, C. and Thomas, M., 2007. SNF1/AMPK/SnRK1 kinases, global regulators at the heart of energy control?. Trends in plant science, 12(1), pp.20-28.

Purcell, P.C., Smith, A.M. and Halford, N.G., 1998. Antisense expression of a sucrose nonfermenting1related protein kinase sequence in potato results in decreased expression of sucrose synthase in tubers and loss of sucroseinducibility of sucrose synthase transcripts in leaves. The Plant Journal,14(2), pp.195-202.

Quinlan, A.R. and Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6), pp.841-842.

R Development Core Team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org (January, 2016).

Radchuk, R., Emery, R.J., Weier, D., Vigeolas, H., Geigenberger, P., Lunn, J.E., Feil, R., Weschke, W. and Weber, H., 2010. Sucrose nonfermenting kinase 1 (SnRK1) coordinates metabolic and hormonal signals during pea cotyledon growth and differentiation. The Plant Journal, 61(2), pp.324-338.

Radchuk, R., Radchuk, V., Weschke, W., Borisjuk, L. and Weber, H., 2006. Repressing the expression of the SUCROSE NONFERMENTING-1-

RELATED PROTEIN KINASE gene in pea embryo causes pleiotropic defects of maturation similar to an abscisic acid-insensitive phenotype. Plant Physiology, 140(1), pp.263-278.

Raji, A.A., Dixon, A.G.O. and Ladeinde, T.A.O., 2007. Agronomic traits and tuber quality attributes of farmer grown cassava (Manihot esculenta) landraces in Nigeria. Journal of Tropical Agriculture, 45, pp.9-13.

Renz, A., Merlo, L. and Stitt, M., 1993. Partial purification from potato tubers of three fructokinases and three hexokinases which show differing organ and developmental specificity. Planta, 190(2), pp.156-165.

Resende, R.T., Resende, M.D.V., Silva, F.F., Azevedo, C.F., Takahashi, E.K., SilvaJunior, O.B. and Grattapaglia, D., 2017. Regional heritability mapping and genomewide association identify loci for complex growth, wood and disease resistance traits in Eucalyptus. New Phytologist, 213(3), pp.1287-1300.

Riggio, V. and Pong-Wong, R., 2014, October. Regional Heritability Mapping to identify loci underlying genetic variation of complex traits. In BMC proceedings (Vol. 8, No. 5, p. 1). BioMed Central.

Riggio, V., Matika, O., Pong-Wong, R., Stear, M.J. and Bishop, S.C., 2013. Genome-wide association and regional heritability mapping to identify loci underlying variation in nematode resistance and body weight in Scottish Blackface lambs. Heredity, 110(5), pp.420-429.

Rolland, F., Moore, B. and Sheen, J., 2002. Sugar sensing and signaling in plants. The plant cell, 14(suppl 1), pp.S185-S205.

SafoKantanka, O. and OwusuNipah, J., 1992. Cassava varietal screening for cooking quality: relationship between dry matter, starch content, meali-

ness and certain microscopic observations of the raw and cooked tuber. Journal of the Science of Food and Agriculture, 60(1), pp.99-104.

Saithong, T., Rongsirikul, O., Kalapanulak, S., Chiewchankaset, P., Siriwat, W., Netrphan, S., Suksangpanomrung, M., Meechai, A. and Cheevadhanarak, S., 2013. Starch biosynthesis in cassava: a genome-based pathway reconstruction and its exploitation in data integration. BMC systems biology, 7(1), p.75.

Shirali, M., Pong-Wong, R., Navarro, P., Knott, S., Hayward, C., Vitart, V., Rudan, I., Campbell, H., Hastie, N.D., Wright, A.F. and Haley, C.S., 2015. Regional heritability mapping method helps explain missing heritability of blood lipid traits in isolated populations. Heredity.

Sorey, J.D. and Tibshirani, R., 2003. Statistical significance for genome-wide experiments. Proc. Natl Acad. Sci. USA, 100, pp.9440-9445.

Ssemakula, G., Dixon, A.G.O. and 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), pp.14-20.

Sweetlove, L.J., Kruger, N.J. and Hill, S.A., 2001. Starch synthesis in transgenic potato tubers with increased 3-phosphoglyceric acid content as a consequence of increased 6-phosphofructokinase activity. Planta, 213(3), pp.478-482.

Tai, H.H., Goyer, C. and Murphy, A.M., 2013. Potato MYB and bHLH transcription factors associated with anthocyanin intensity and common scab resistance. Botany, 91(10), pp.722-730.

Takatsuji, H., 1998. Zinc-finger transcription factors in plants. Cellular and Molecular Life Sciences CMLS, 54(6), pp.582-596.

Takatsuji, H., 1999. Zinc-finger proteins: the classical zinc finger emerges in contemporary plant science. Plant molecular biology, 39(6), pp.1073-1078.

Tanaka, M., Takahata, Y., Nakayama, H., Nakatani, M. and Tahara, M., 2009. Altered carbohydrate metabolism in the storage roots of sweetpotato plants overexpressing the SRF1 gene, which encodes a Dof zinc finger transcription factor. Planta, 230(4), pp.737-746.

Tang, X., Su, T., Han, M., Wei, L., Wang, W., Yu, Z., Xue, Y., Wei, H., Du, Y., Greiner, S. and Rausch, T., 2016. Suppression of extracellular invertase inhibitor gene expression improves seed weight in soybean (Glycine max). Journal of Experimental Botany, p.erw425.

Tiessen, A., Prescha, K., Branscheid, A., Palacios, N., McKibbin, R., Halford, N.G. and Geigenberger, P., 2003. Evidence that SNF1related kinase and hexokinase are involved in separate sugarsignalling pathways modulating posttranslational redox activation of ADPglucose pyrophosphorylase in potato tubers. The Plant Journal, 35(4), pp.490-500.

Toledo-Ortiz, G., Huq, E. and Rodrguez-Concepcin, M., 2010. Direct regulation of phytoene synthase gene expression and carotenoid biosynthesis by phytochrome-interacting factors. Proceedings of the National Academy of Sciences, 107(25), pp.11626-11631.

Uemoto, Y., Pong-Wong, R., Navarro, P., Vitart, V., Hayward, C., Wilson, J.F., Rudan, I., Campbell, H., Hastie, N.D., Wright, A.F. and Haley, C.S., 2013. The power of regional heritability analysis for rare and common variant detection: simulations and application to eye biometrical traits. Frontiers in genetics, 4, p.232.

VanRaden PM: Efficient methods to compute genomic predictions. J Dairy Sci. 2008, 91: 4414-4423. 10.3168/jds.2007-0980.

Vimala, B., Nambisan, B., Theshara, R. and Unnikrishnam, M., 2008. Variability of carotenoids in yellow-flesh cassava (Manihot esculatar Crantz).Geneconserve. Pro. br, 1. Wang, E., Wang, J., Zhu, X., Hao, W., Wang, L., Li, Q., Zhang, L., He, W., Lu, B., Lin, H. and Ma, H., 2008. Control of rice grain-filling and yield by a gene with a potential signature of domestication. Nature genetics, 40(11), pp.1370-1374.

Wang, L. and Ruan, Y.L., 2012. New insights into roles of cell wall invertase in early seed development revealed by comprehensive spatial and temporal expression patterns of GhCWIN1 in cotton. Plant physiology, 160(2), pp.777-787.

Weschke, W., Panitz, R., Gubatz, S., Wang, Q., Radchuk, R., Weber, H. and Wobus, U., 2003. The role of invertases and hexose transporters in controlling sugar ratios in maternal and filial tissues of barley caryopses during early development. The Plant Journal, 33(2), pp.395-411.

Weschke, W., Panitz, R., Sauer, N., Wang, Q., Neubohn, B., Weber, H. and Wobus, U., 2000. Sucrose transport into barley seeds: molecular characterization of two transporters and implications for seed development and starch accumulation. The Plant Journal, 21(5), pp.455-467.

Wolfe, M.D., Rabbi, I.Y., Egesi, C., Hamblin, M., Kawuki, R., Kulakow, P., Lozano, R., Carpio, D.P.D., Ramu, P. and Jannink, J.L., 2016. Genome-wide association and prediction reveals genetic architecture of cassava mosaic disease resistance and prospects for rapid genetic improvement. The plant genome, 9(2).

Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J.A., Kutalik, Z. and Amin, N., 2014. Defining

the role of common variation in the genomic and biological architecture of adult human height. Nature genetics, 46(11), pp.1173-1186.

Xue-Fei, D., Na, C., Li, W., Xiao-Cui, Z., Bo, Q., Tian-Lai, L. and Guo-Liang, Z., 2012. The SnRK Protein Kinase Family and the Function of SnRK1 Protein Kinase. International Journal of Agriculture and Biology, 14(4).

Yang, J., An, D. and Zhang, P., 2011. Expression profiling of cassava storage roots reveals an active process of glycolysis/gluconeogenesis. Journal of integrative plant biology, 53(3), pp.193-211.

Zanor, M.I., Osorio, S., Nunes-Nesi, A., Carrari, F., Lohse, M., Usadel, B., Khn, C., Bleiss, W., Giavalisco, P., Willmitzer, L. and Sulpice, R., 2009. RNA interference of LIN5 in tomato confirms its role in controlling Brix content, uncovers the influence of sugars on the levels of fruit hormones, and demonstrates the importance of sucrose cleavage for normal fruit development and fertility. Plant physiology, 150(3), pp.1204-1218.

Zeng, Y., Navarro, P., Fernandez-Pujals, A.M., Hall, L.S., Clarke, T.K., Thomson, P.A., Smith, B.H., Hocking, L.J., Padmanabhan, S., Hayward, C. and MacIntyre, D.J., 2016. A Combined Pathway and Regional Heritability Analysis Indicates NETRIN1 Pathway is Associated with Major Depressive Disorder. Biological Psychiatry.

Zhang, Y., Jung, C.S. and De Jong, W.S., 2009. Genetic analysis of pigmented tuber flesh in potato. Theoretical and Applied Genetics, 119(1), pp.143-150.

CHAPTER 4

REGIONAL CO-HERITABILITY MAPPING (RCHM) PROVIDES INSIGHTS INTO THE CO-INHERITANCE PATTERNS OF DRY MATTER (DM) CONTENT, ROOT COLOR AND FRESH ROOT YIELD (FYLD) IN DIFFERENT SUBPOPULATIONS OF AFRICAN CASSAVA.

4.1 Abstract:

We attempted to gain insights into the co-inheritance of pairs of traits; dry matter (DM) content and fresh yield (FYLD), or DM and root yellowness (B). These are core cassava traits representing two specialized products, high starch white and provitamin-A yellow cassava respectively. Consequently, we developed the Regional co-heritability mapping (RcHM) procedure. The RcHM started by estimating SNP effects from a bivariate mixed model. Then the genome was fragmented into 54 segments of approx. 18Mb. SNP effects were then summed per segment to obtain genomic segment values (GSVs) of clones for each trait. Subsequently, genomic segment correlations (GSCs) signifying co-inheritance were estimated as pairwise correlations between trait GSVs. The GSC significance was assessed via resampled residual bootstrapping. White cassava GSCs were mostly (34/54) favorable (positive) for DM and B with 15% significant. Yellow cassava GSCs were mostly (45/54) unfavorable (30% significant). The yellow-plus-white had mostly (34/54) unfavorable GSCs (39% significant). For specialized starch trials harvested at 12 months after planting (MAP) - HS1 and 14 MAP (HS2); HS1 GSCs for DM and FYLD were mostly (43/54) favorable (52% significant). HS2 had fewer favorable (38/54) GSCs than HS1 (20% significant). Genomic correlations were 0.14, -0.30 and -0.08 for white, yellow

and yellow-plus-white respectively; those for HS1 and HS2 were 0.09 and 0.12. These demonstrate good potential for developing high starch white cassava and a limited prospect for yellow cassava development. RcHM was useful for mapping the co-inheritance of complex traits in cassava.

4.2 Background:

Breeding for the micronutrient biofortification of world staples has been recognized as a practical technique to combat hidden hunger (Tanumihardjo et al., 2008; Bouis et al., 2010; Bouis et al., 2011). The HarvestPlus program of the International Institute of Tropical Agriculture (IITA) for cassava breeding in Nigeria targets fortifying cassava with provitamin A (a vitamin A precursor) to provide consumers in Africa with a daily requirement of vitamin A in their diet (Sayre et al., 2011; Bouis, 2014). To meet this target, African cassava breeders develop so called yellow cassava varieties (Bouis, 2014). The yellowness of the cassava root is an indication of the total carotenoid content of which beta-carotene (provitamin A) is a member (Iglesias et al., 1997; Maziya-Dixon et al., 2010; Ceballos et al., 2013). Therefore the fast and easy to measure yellow flesh coloration of the cassava root is an indicator trait for beta-carotene (Iglesias et al., 1997; Ceballos et al., 2013). However, the value of cassava is the percentage of dry matter (DM) content in the total fresh root yield (FYLD) per hectare of this crop (Kawano et al., 1987). This percentage includes mainly carbohydrates (90%) which provide calories (Kawano et al., 1987). It is vital for the HarvestPlus program to develop yellow cassava varieties with high DM content (Bouis, 2014).

African cassava breeders have previously identified a negative correlation between DM and root yellowness in the African cassava germplasm which has posed a challenge to this biofortification target (Akinwale et al., 2010; Esuma et al., 2016). The consequence of this is that most of the yellow cassava varieties have low DM content making them economically unattractive for farmers and processors even though consumer adoption is promising (Bouis, 2014; Oparinde et al., 2016). However this negative correlation has not been an issue in South American cassava germplasm. Breeders at the International Center for Tropical Agriculture (CIAT) in Colombia have reported concurrent improvement of both traits via rapid recurrent selection (Ceballos et al., 2013). Therefore it is useful to understand the genetic basis for this observed genetic correlation between DM and root yellowness in the African cassava germplasm in order to develop varieties that carry both traits.

One way to understand the basis for this observed negative genetic correlation between DM and root flesh yellowness is by mapping the genetic correlation of genomic segments across the cassava genome. In this map, a favourable genomic segment will be defined as a segment on the chromosome that has correlation between two traits in a desired direction. In our case with DM and root yellowness, a segment with a positive genetic correlation is favorable. This genome-wide map showing the distribution of favourable and unfavorable genomic segments will reveal the co-inheritance profile of these traits. Genomic segments may be useful for designing a breeding strategy that selects on favourable segments (Daetwyler et al., 2015). Selection based on genomic estimated breeding values (GEBVs) gives the highest response for the next generation (Cole and VanRaden, 2011; Kemper et al., 2012). However, antagonistic relationships among loci that affect multiple traits under selection may limit the

GEBV of a trait (Cole and VanRaden, 2011). These antagonistic relationships may be based on biological mechanisms underlying these traits for instance antagonistic pleiotropic loci or the occurrence of favourable and unfavourable alleles on the same haplotype due to linkage thereby yielding a negative correlation between two traits (Cole and VanRaden, 2011; Carter and Nguyen, 2011). To use information from genomic segments, selection may proceed by making the effects of antagonistic genomic segments neutral in the selection index.

To understand the co-inheritance of DM and root flesh yellowness using a genome-wide correlation map of genomic segments, we developed a method termed the Regional co-Heritability Mapping (RcHM). The RcHM is based on a bivariate mixed model and captures the genetic correlation of traits of interest when estimating the marker effects for these traits. These marker effects are subsequently used to obtain genomic segment values (GSVs) which are used for developing genome-wide segment correlation maps. Genomic segments may harbour adjacent polymorphisms in a block that are all associated with traits of interest (Kemper et al., 2014). We presume that if the combined function of these polymorphisms are favourable for a combination of traits, then this genomic segment may be of interest to a breeder. In order to understand the value of a group of adjacent polymorphisms in the genome for a combination of traits and how these influence the co-inheritance of these traits, a genome-wide segment correlation map may be useful. These correlation maps reveal the relationship between two traits of interest at specific genomic segments and provide information to the breeder useful for managing traits that are unfavourably associated in a breeding program, as in the case of DM and root yellowness in the African cassava germplasm. Such decisions may involve the use of gene editing strategies like CRISPR (Ma et al., 2016) or the use of a random mating pro-

gram without selections using donor germplasms to breakup favourable and unfavourable loci which are in repulsion phase.

The objectives of this paper are:

1. To understand the genetic basis of the negative correlation between the DM and root yellowness in the African cassava germplasm using genomic segment correlation maps and how this influences development of high DM yellow varieties.
2. To understand the co-inheritance patterns of DM and FYLD in specialized high starch trials of high DM white clones and how these influence development of high DM white cassava varieties.
3. To understand the sensitivity of genomic segment correlations to changes in whole genome correlations between DM and root yellowness.

4.3 Materials and Methods:

4.3.1 Cassava data:

Phenotypic data used in our analysis were from Genetic Gain (GG) trials conducted by the cassava breeding program at the Institute of Tropical Agriculture (IITA), Ibadan, Nigeria from 2013 to 2015. The GG population is a breeding population developed from the 1970s to 2007 at the IITA (Maziya-Dixon et al., 2007; Okechukwu and Dixon, 2008; Ly et al., 2013). For the analysis in this study, we used GG trials planted in an augmented design. The design consisted of a layout of between 18 to 30 blocks with 22 accessions and two checks in each

block. Accession plots were a single row (1m x 1m spacing) of five-plant stands without borders which were mostly unreplicated. These trials were conducted in three locations in Nigeria including Ibadan (7.40 N, 3.90 E), Mokwa (9.3 N, 5.0 E), and Ubiaja (6.66 N, 6.38 E). Some core agronomic traits were measured for these trials but we concentrate here on percentage dry matter (DM) of storage roots, which measures root dry weight as the percentage of the root fresh weight, fresh weight of harvested roots expressed in tons per hectare (FYLD), pulp color (PLPCOL) a binary trait rated on a scale of 1 indicating white to light cream flesh root, or 2 indicating deep cream to yellow flesh root, and root flesh color measured using a KONICA MINOLTA CR-400 series chromameter which captures the RGB color space in L, A and B units (Leon et al., 2006). Our focus was on B which captures blue as negative values and yellow as positive values (Broadbent, 2004; Leon et al., 2006). The oven method was used for DM: 100g grated root sample (with thorough mixing of 10-15 randomly selected roots from a plot) were collected per accession, oven dried, and DM is expressed as the residual weight. We further divided the GG population (713 clones) into two subpopulations of white (451 clones) and yellow (262 clones) cassava using the PLPCOL score or 1 and 2 for the white and yellow populations, respectively. The genotypic data used in this study was described in Okeke et al. 2017.

4.3.2 Specialized starch trials data:

Specialized trials were conducted in collaboration between the IITA and the National Root Crops Research Institute (NRCRI), Umudike, Nigeria from 2012 to 2014. The aim of these trials was to select high yielding and high starch cassava clones for the starch industry. These trials consisted of 52 clones planted

in a randomized complete block (RCB) design with 3 to 4 replications. Plots were six rows (1m x 0.8m spacing) of 8 plant stands without borders. These trials were conducted in 15 locations in Nigeria including Abuja (9.06 N, 7.40 E), Akure (7.26 N, 5.19 E), Ikenne (6.88 N, 3.70 E), Ilorin (8.48 N, 4.55 E), Mokwa (9.30 N, 5.0 E), Ubiaja (6.66 N, 6.38 E), Onne (4.74 N, 7.04 E), Warri (5.56 N, 5.79 E), Zaria (11.30 N, 7.69 E), Akwa-Ibom (5.07 N, 7.89 E), Benue (7.58 N, 8.69 E), Calabar (4.98 N, 8.34 E), Imo (5.52 N, 7.11 E), Taraba (8.71 N, 10.97 E) and Umudike (5.47 N, 7.54 E). As before, some core agronomic traits were measured for these trials but we concentrate here on the traits DM and FYLD. These trials were planted on September 2012 and 2013 but in order to understand the effect of time of harvest on DM and FYLD, they were harvested in two sets of 16 plants per plot each at 12 months after planting (referred to as HS1) and 14 months after planting (HS2). Most of these 52 clones were members of the GG population. The harvests had 2,113 records for HS1 and 1,797 for HS2.

4.3.3 Data analysis:

Regional co-heritability mapping (RcHM):

RcHM was carried out as follows:

1. The following bivariate linear mixed model was fit using whole genome SNP markers:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e} \\ \mathbf{y} = (\mathbf{y}_{DM}^{'}, \mathbf{y}_{B}^{'})^{'} \quad u = (\mathbf{u}_{DM}^{'}, \mathbf{u}_{B}^{'})^{'}; \quad \mathbf{e} &= (\mathbf{e}_{DM}^{'}, \mathbf{e}_{B}^{'})^{'} \end{aligned} \tag{4.1}$$

   where $\mathbf{y}$ is a concatenated vector for traits DM and B recorded for n clones,

$\mathbf{X}$ and $\mathbf{Z}$ are block diagonal design matrices represented as $\mathbf{diag}(\mathbf{X_{DM}}, \mathbf{X_B})$ and $\mathbf{diag}(\mathbf{Z_{DM}}, \mathbf{Z_B})$, respectively allowing for missing clones and observations. $\mathbf{X}$ is the design matrix for fixed effects β (with components for the grand mean, Location-Year-Trial nested effects, a Rep effect for the specialized trials and 10 principal component vectors from decomposition of the $\mathbf{K}$ matrix) and $\mathbf{Z}$ is the design matrix for random genomic effects $\mathbf{u}$. The marginal density of y is multivariate normal (Nm):

$$\begin{aligned} (\mathbf{y}\,|\,\beta, \mathbf{R}, \mathbf{G}) &\sim N_m(\mathbf{X}\beta, \mathbf{V}) \\ \mathbf{V} = \mathbf{Z}(\mathbf{G} \otimes \mathbf{K})\mathbf{Z}^T + \mathbf{R} \otimes \mathbf{I}); \quad \hat{\mathbf{u}} &= (\mathbf{G} \otimes \mathbf{K})\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta) \end{aligned} \tag{4.2}$$

where $\mathbf{G}$ and $\mathbf{R}$ are 2×2 symmetric genomic and error covariance matrices respectively, $\mathbf{K}$ is an additive genomic relationship matrix for n clones generated from SNP markers the first method in VanRaden, (2008), $\mathbf{I}$ is an identity matrix and $\mathbf{u}$ is a vector of clonal genomic estimated breeding values (GEBVs) and for the DM and B traits. Estimation of the parameters in model (5.1) were performed using the REML procedure implemented in the airemlf90 program (Masuda et al., 2015) from which BLUEs of fixed effects and BLUPs of random effects were obtained by solving the mixed model equations (MME; Henderson, 1973).

2. Following procedure (1), SNP effects were then calculated for both DM and B as:

$$\hat{\mathbf{g}}_\mathbf{t} = \lambda \mathbf{D}\mathbf{M}^T\mathbf{K}^{-1}\hat{\mathbf{u}}_\mathbf{t} \tag{4.3}$$

where $\hat{g}_t$ is a vector of SNP effects for trait t (DM or B), $\mathbf{M}$ is a SNP matrix first coded -1, 0, 1 for reference homozygote, heterozygote, and alternate homozygote, then column centered, K is the same as in (1) above and $\hat{u}_t$ is a vector of clonal genomic effect for the trait t. λ is a ratio of variances

or normalizing constant according to VanRaden (2008) and **D** is a diagonal matrix of weights of SNP variances. For this study the weights of SNP were not used, so $\mathbf{D} = \mathbf{I}$ (identity matrix). These SNP effects were generated using the postGSf90 program (Aguilar et al., 2014).

3. Each chromosome was then divided into three windows containing equal numbers of SNPs. These windows are what we refer to as genomic segments. Windows ranged from 1,577 SNPs on chromosome 7 to 3,242 SNPs on chromosome 1. The average size of segments across chromosomes was 18Mb with a total of 54 segments genome-wide.

4. Subsequently, genomic segment values were calculated as (Koivula et al., 2012):

$$\mathbf{G\hat{S}V_t} = \mathbf{M}_{seg}^{T}\mathbf{\hat{g}_t} \tag{4.4}$$

for all segments in the genome where t can be either DM or B, $\mathbf{M}_{seg}$ was a centered marker matrix for the SNPs in a segment and $\mathbf{\hat{g}_t}$ were the SNP effects of these segment SNPs. Matrix computations were done in R (R Core Team, 2017).

5. Genomic correlation for DM and B at each segment were then calculated as the pearson correlation across cassava clones between the $\mathbf{GSV}_{DM}$ and $\mathbf{GSV}_{B}$.

We carried out the RcHM analysis for white and yellow subpopulations and for yellow-plus-white combined. In the case of the specialized starch trials, we carried out RcHM analysis for the two harvest sets HS1 and HS2 replacing B with FYLD. For reasons due to population structure, 10 principal component

vectors were added to the fixed effects. We discuss the rationale behind this later in the discussion section.

4.3.4 Resampled residual bootstrap analysis for assessing significance:

In order to assess the significance of genomic segment correlations in the genome-wide correlation map, we carried out a bootstrap analysis based on residual resampling (Douc and Capp, 2005) using the following procedure:

1. Fit the bivariate model as in (1) from the RcHM procedure above. Extract residuals. Then new phenotypes (y*) for DM and B (or FYLD as appropriate) were obtained by adding resampled residuals to the original phenotypes: $y^* = y + \hat{e}$.
2. RcHM analysis as outlined above was carried out using new phenotypes y^* 500 times to obtain sampling distributions for genetic (co)variance parameters from the bivariate model as well as distributions for genomic segment correlations.
3. Significance threshold for genomic segment correlations were set at 5% corresponding to the 2.5% and 97.5% quantiles of the sampling distributions for each of the genomic segment correlations from procedure (2) above. These thresholds were also the 95% confidence intervals.
4. Significance was determined by a genomic segment correlation confidence interval that did not overlap with zero.

4.3.5 Sensitivity analysis:

To understand the impact of changing genomic correlations on segment correlations, we carried out the RcHM analysis for DM and B as described above by fixing genetic covariances corresponding to genetic correlations of 0.5, 0 and -0.5 while genetic variances, error variances and covariances remained unchanged. Sensitivity analysis commenced as follows:

1. Recall that generating GEBVs from Model (4.2) is as follows:

$$\hat{\mathbf{u}} = (\mathbf{G} \otimes \mathbf{K})\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta)$$
$$\mathbf{G} = \begin{bmatrix} \sigma^2_{DM} & X\sigma_{DM}\sigma_B \\ X\sigma_{DM}\sigma_B & \sigma^2_B \end{bmatrix} \tag{4.5}$$

2. We fixed X in $X\sigma_{DM}\sigma_B$ to -0.5, 0 and 0.5 to generate $\hat{\mathbf{u}}_\mathbf{t}$ for estimating SNP effects in Equation (4.3) in the RcHM analysis while σ^2_{DM} and σ^2_B remained same as in the original analysis.
3. Subsequently, procedures 2 – 5 were carried out as described in the RcHM analysis above.

4.4 Results:

4.4.1 Co-inheritance of DM and B based on RcHM analysis for cassava subpopulations:

The estimated whole genome variances and correlations (Table 4.2) for DM and B in the white, yellow and combined cassava subpopulations provide context

to the segment correlations (Figures 5.1, 5.2, and 5.3 for white, yellow, and combined, respectively). For the white cassava subpopulation, we observed that 38 out of 54 genomic segments showed a moderate to strong positive genetic correlation between DM and B with 15% (8/54) of these significantly differing from 0 (Figure 4.1). Genomic segment correlations ranged from -0.74 to 0.96. For the yellow cassava subpopulation, we observed that segments in the genome overwhelmingly showed a moderate to strong negative genetic correlation between DM and B with 30% (16/54) of these segments having correlations significantly differing from 0 (Figure 4.2). Genomic segment correlations ranged from -0.93 to 0.62. Lastly, in the combined white and yellow cassava subpopulation, we observed that segments had moderately negative or positive genetic correlations with 65% of these segments being negatively correlated (Figure 4.3). Genomic segment correlations ranged from -0.90 to 0.91. However, 39% (21/54) of these genomic segments had correlations that differed significantly from 0.

4.4.2 Co-inheritance of DM and FYLD based on RcHM analysis and effects of the time of harvest:

For the specialized starch trials involving white clones harvested at different ages, we also observed varied genome-wide correlation patterns for the traits DM and FYLD. For HS1 harvested at 12 months after planting (MAP), we observed that genomic segments had both moderate to strongly negative and positive genetic correlations with 79% of the segments correlations being positive (Figure 4.4). Genomic segment correlations ranged from -0.82 to 0.94, of which 52% (28/54) were significantly different from 0 (Figure 4.4). For HS2 harvested

14 MAP, we also observed moderate to strong segment genetic correlations with 69% of these segments being positively correlated (Figure 4.5). Also 20% (11/54) of these segment correlations were significantly different from 0. Genomic segment correlations ranged from -0.94 to 0.94. HS1 had more favourable segment associations with 25 significant positively correlated segments compared to HS2 with 10. These indicate benefits for harvesting at 12 MAP.

4.4.3 Sensitivity analysis:

We observed that segment correlations were sensitive to changes in the genomic correlation between DM and B. Genomic segment correlations were markedly different for all populations when the genomic correlations changed from -0.5 to 0 or from 0 to 0.5. For the analysis with 0.5 genomic correlation, we observed 94% positive segment correlations for white cassava; a mix of segment correlations for yellow with 56% positive and 31% negative (Figures 5.6 and 5.7). The yellow-plus-white also had a mix of segment correlations with 67% positive and 28% negative (Figure 4.8). For the analysis with zero genomic correlation, we observed a mix of segment correlations for white cassava with 57% being positive and 37% negative. Segment correlations were mostly (63%) negative for the yellow, while 35% were positive for the yellow-plus-white population with 54% being negative. Segment correlations were moderate to high for all populations. For the analysis with -0.5 genomic correlation, 7% segment correlations were positive and 93% negative for white cassava; 6% segment correlations were positive for the yellow and 90% being negative while 19% segment correlations were positive for the yellow-plus-white population with 76% negative (Figures 5.6, 5.7 and 5.8). Again, segment correlations were moderate to

high for all populations.

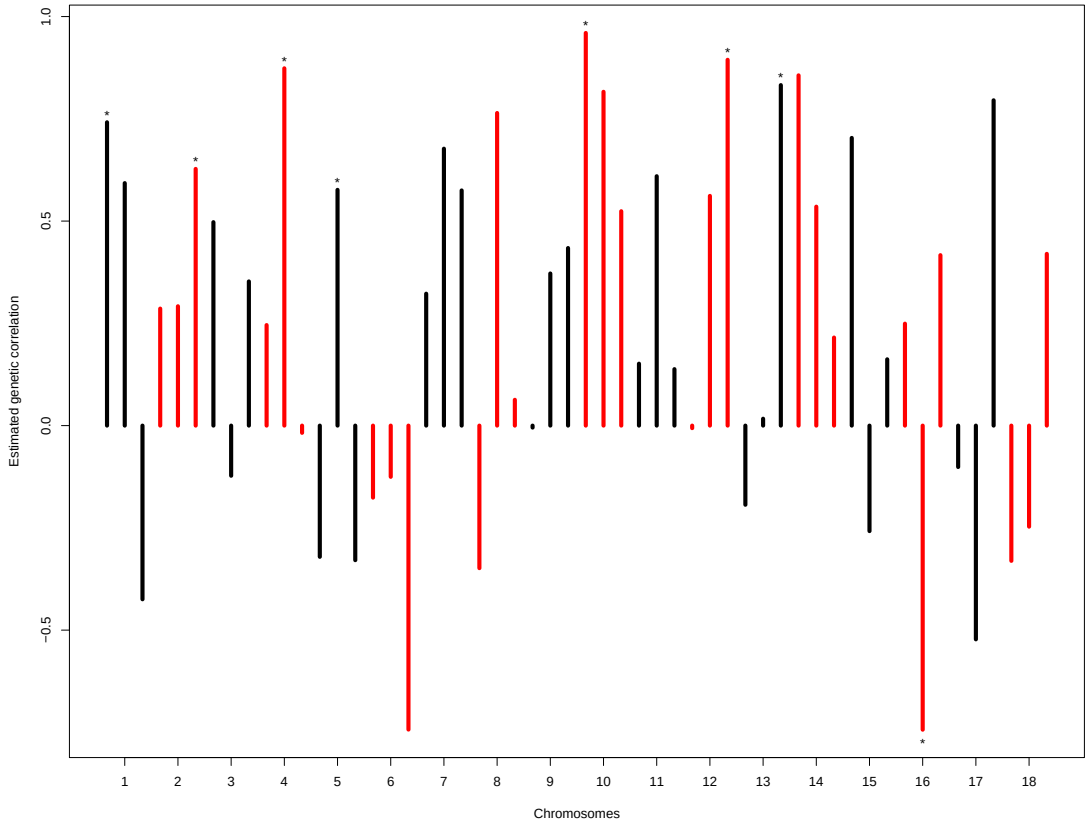

Figure 4.1: **Genomic segment correlation map between DM and B for white cassava.**

The (*) indicate significance of correlations at 5% level.

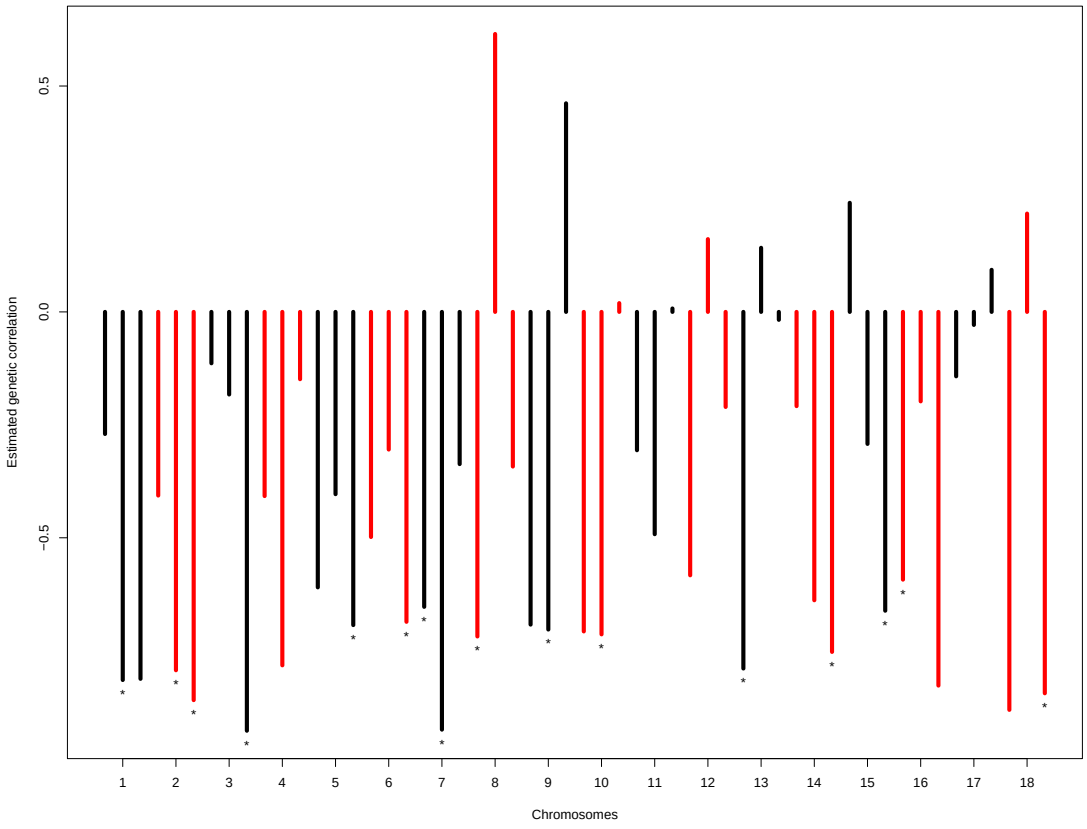

Figure 4.2: **Genomic segment correlation map between DM and B for yellow cassava.**

The (*) characters are as in Figure 4.1.

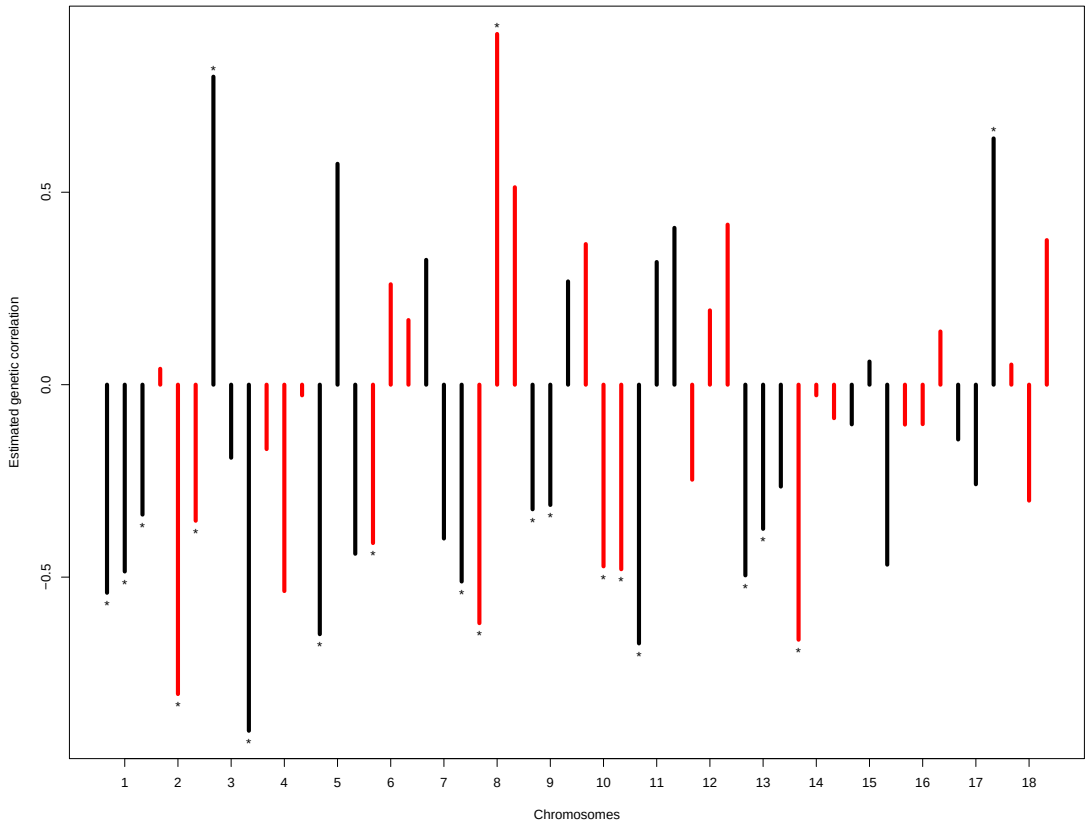

Figure 4.3: **Genomic segment correlation map between DM and B for yellow-plus-white cassava.**

The (*) characters are as in Figure 4.1.

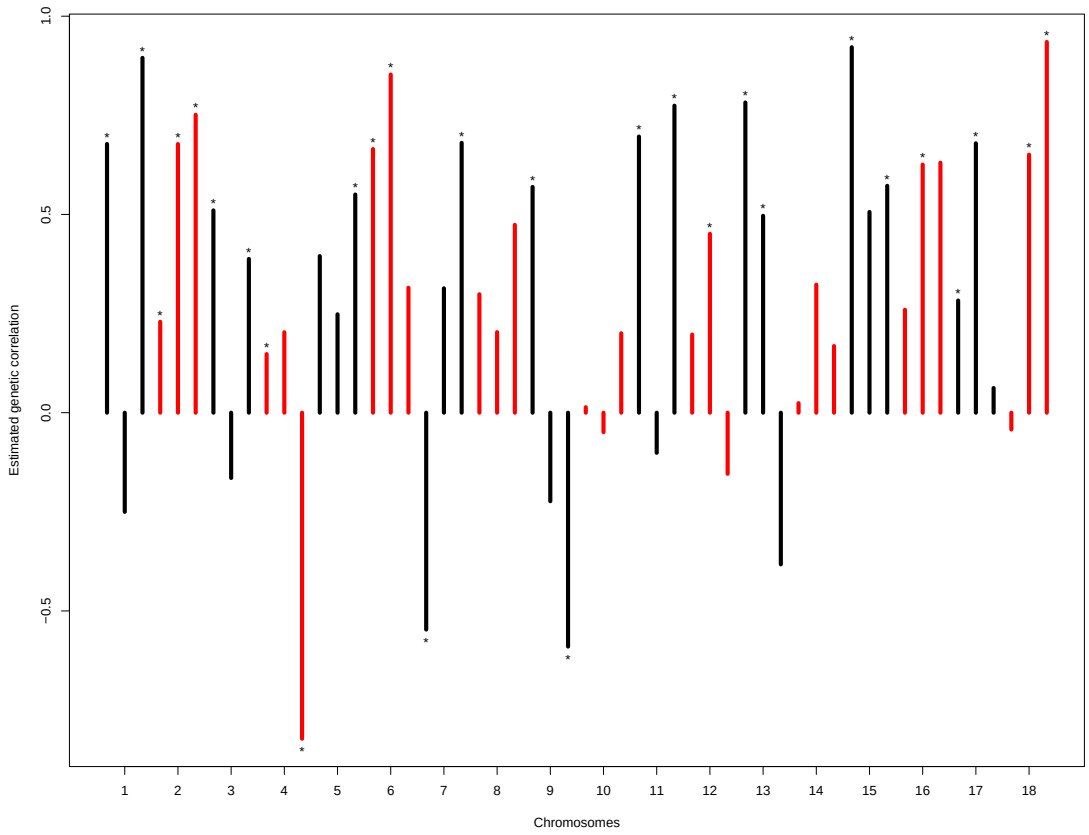

Figure 4.4: **Genomic segments correlation map between DM and FYLD for white cassava harvested 12 MAP.**

The (*) indicate significance at 5% level.

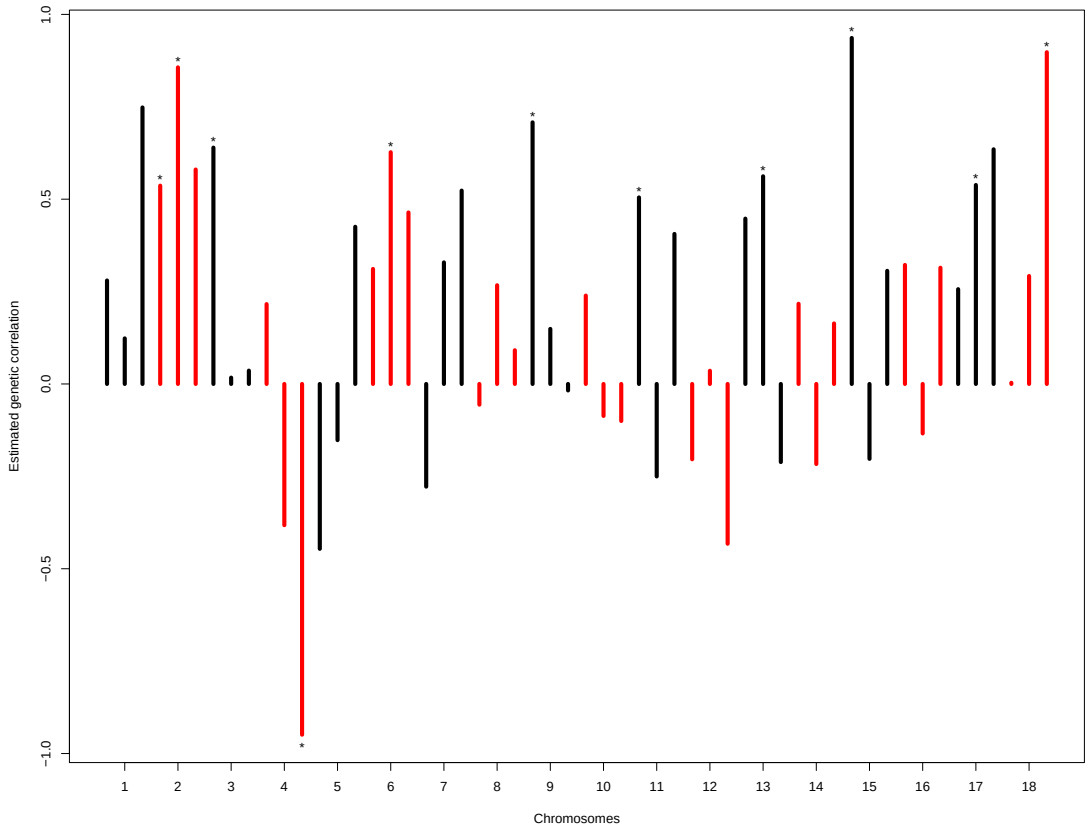

Figure 4.5: **Genomic segment correlation map between DM and FYLD for white cassava harvested 14 MAP.**

The (*) indicate significance at 5% level.

4.5 Discussion:

We have attempted to gain insights into the co-inheritance of pairs of traits DM and B, or DM and FYLD (core and valuable traits in cassava breeding) using the RcHM analysis. The RcHM is primarily based on a random SNP effects bivariate model that captures both strong and moderate SNP-trait associations (Frszczak and Szyda, 2016) and also accounts for genetic correlation between traits thus obtaining more accurate SNP effects for both traits (Ferreira et al., 2009; Galesloot et al., 2014). To understand the impact of co-inherited genomic blocks (that may harbour linked or pleiotropic variants) on traits, the RcHM estimates the genomic correlations from GSVs of both traits which are a reflection of the effect of these segments on these traits. If a segment has a moderate to strong correlation in the direction that is of interest to the breeder, then it is a favourable segment and may be exploited for improvement of both traits. This exploitation may proceed first by estimating combined segment values (CSVs) for a trait as: $CSV = \Sigma_{i=1}^{n} GSV_i \alpha_i$ where i goes from 1 to n, the total number of segments in the genome and α_i is the weight given to the i^{th} GSV. If $\alpha_i = 1$ for all segments in the genome, then the CSV equals the GEBV. α_i represents importance of a segment based on whether the segment is favourable or not. CSVs can be used in place of GEBVs in a selection index. Segment correlations can also be visualized in a plot (termed correlation maps) to understand the co-inheritance pattern of two traits and to help inform the breeder when making decisions in a breeding program. In this study, we used these genome-wide correlation maps to understand the differences in DM and B co-inheritance in 3 subpopulations of the African cassava and for DM and FYLD for two harvest times in specialized starch trials for high DM white cassava.

4.5.1 Developing high DM white cassava varieties:

The genome-wide correlation maps from the RcHM analysis for the white cassava subpopulation show good prospects for developing white clones with high DM content (Figure 4.1). We observed that most (70%) of the genomic segments were favourable with positive genomic correlations. An index based on multitrait breeding values will be beneficial for making gains on these traits (Bauer and Lon, 2008). The overwhelming agreement on the positive direction for DM and B by genomic segment correlations in white cassava shows that rapid progress can be achieved on these traits based on multitrait genomic selection (GS). The genomic correlation for DM and B from the bivariate GBLUP model for white cassava was moderate (0.14) further supporting that gains can be made on selection due to correlated response. However, we recall that more value is achieved for high DM white clones with higher FYLD (Kawano et al., 1987). We tried to understand the plausibility of developing high yielding and high DM white clones using genome-wide correlation maps from the RcHM analysis based on data from specialized starch trials conducted at different agroecological zones in Nigeria. We know (Hammer et al., 1987; Ebah-Djedji et al., 2012) that the time of harvest influences the DM content in cassava roots with more DM content observed when harvest is delayed till the onset of the dry season. However, the genome-wide correlation maps showed complex co-inheritance patterns for DM and FYLD for the different harvest times HS1 (Figure 4.4), HS2 (Figure 4.5). We observed mostly (79%) favourable genomic segments in HS1 while HS2 had less (69%) favourable segments. These results suggest that harvesting 12 MAP may be beneficial for high DM and high yielding cassava production. Also, the development of high yielding and high DM white clones is plausible using an index with multitrait genomic breeding val-

ues. Such an index has been shown to produce more gains in self pollinating crops (Bauer and Lon, 2008). Based on the HS1 map, 20% of genomic segment correlations for DM and FYLD at 12 months harvest were antagonistic. This leads to a situation where the FYLD CSVs may be explored for use in this multitrait index. However more studies are required to understand how CSVs may be utilized in a breeding program. Another interesting view of segment correlations between HS1 and HS2 (Figure 4.11) is the consistently in the number of favourable (50%) or unfavourable segments (22%). However, the differences are due to favourable segments in HS1 from chromosomes 4, 5, 8, 10, 12, 15 and 16 or those in HS2 from chromosomes 1, 2, and 17 (Figure 4.11). These may represent opposite biological processes that affect DM and yield due to age of plants or time of harvest. Interestingly, we observed 31% (17/54) consistent favourable segments between HS1 and white cassava (Figures 5.12). These represent genomic links between DM, B and FYLD in white cassava and may also point to biological processes associated with these traits. These 17 favourable segments further reveal good prospects for developing high DM and high yielding white cassava. Further investigation is required to unravel these.

We further suggest that the white cassava germplasm be maintained as a separate breeding program tasked for developing white varieties. This is because maintaining both white and yellow germplasms as a unit may not be beneficial as seen in Figure 4.3 that shows complicated co-inheritance patterns for DM and B. Mixing these may not encourage defined market products for white and yellow cassava.

4.5.2 Developing high DM yellow cassava varieties:

The genome-wide correlation map for the yellow cassava showed an unfavourable genome-wide co-inheritance profile (Figure 4.2) for the development of yellow clones with high DM content. We observed overwhelmingly (81%) moderate to strong negative segment correlations for which 30% were significant. This may indicate moderate to strong genome-wide antagonistic effects for DM and B for yellow cassava clones. The genomic correlation for DM and B from the bivariate GBLUP model was moderately strong (-0.3) further reflecting a complex challenge for yellow cassava breeding. We argue that in this case, the genome-wide correlation map showed to a large extent how challenging the task of developing high DM yellow varieties can be thus helping to move towards decisions based on innovative strategies to tackle this challenge. It is interesting to recall that this negative genetic correlation between DM and yellowness is not observed in the South American cassava germplasm hence CIAT has been developing high DM yellow cassava over the years via a recurrent selection approach (Ceballos et al., 2013). Also several projects over the years have been hybridizing CIAT germplasms to their African counterparts but this problem persists after years of this introgression (Akinwale et al., 2010, Bouis, 2014). It is difficult to point to either pleiotropy or linkage (in repulsion phase) as the culprit here but an antagonistic genetic correlation resulting from pleiotropic effects is a more challenging situation which can be overcome when an outlier genotype that harbours an alternative biochemical pathway(s) or physiological processes controlled by other genes is found (Luby and Shaw, 2009). In the case of repulsion phase linkage, a recombinant that breaks this correlation is sought after (Esch et al., 2007). However other technologies involving transgenics and gene editing can be pursued when genes involved in the biochemical processes

of DM and B are clearly understood. We also recommend that yellow cassava germplasm be maintained as a separate breeding program and not mixed with the white. This will facilitate the search for recombinants and further lead to the accumulation of valuable data for further analysis. We also encourage random mating of yellow cassava germplasm from IITA and CIAT for multiple generations without selection in a neutral location (outside Africa) to encourage recombination and to remove the influence of the cassava mosaic disease (CMD) which has been posing challenges to exotic yellow clones brought into Africa (Personal communication, Dr. Peter Kulakow and Teddy Hanmakyugh).

Choice of 10PCs for the RcHM analysis:

The 10 principal component (PCs) vectors in the RcHM analysis served to control for population structure across the genome. Population structure may drive genetic correlation between traits in a population like ours (white and yellow cassava) that has diverged in recent past due to drift and selection. Consequently, allele frequencies of loci not related to our traits of interest (DM, B and FYLD) may have also diverged. When this structure is not accounted for, this divergence-induced differences in allele frequencies will result in spurious associations in the RcHM analyses leading to a false positive genetic correlation between the analyzed traits. Fixing PCs in the RcHM analysis served to absorb these population structure effects (Patterson et al., 2006) thus helping in the identification of signals unique to each genomic segment.

The decision for using 10 PCs as fixed effects in the RcHM analysis to correct for population structure were made after performing regression analyses using 10, 15 and 50 PCs onto DM on the white and yellow cassava germplasms (Table 4.1). We found that 10 PCs explained 9% and 24% of the total DM variation in

	White	Yellow
(Intercept)	30.61***	25.11***
	(0.14)	(0.23)
Zpc1	0.30	−20.45***
	(2.67)	(3.36)
Zpc2	17.07***	−32.75***
	(2.86)	(3.34)
Zpc3	17.49***	9.17**
	(2.88)	(3.34)
Zpc4	7.60**	21.20***
	(2.88)	(3.35)
Zpc5	1.44	−6.00
	(2.87)	(3.39)
Zpc6	20.38***	−0.63
	(2.90)	(3.33)
Zpc7	3.26	7.45*
	(2.87)	(3.33)
Zpc8	1.50	9.72**
	(2.94)	(3.34)
Zpc9	8.08**	−5.28
	(2.84)	(3.32)
Zpc10	−3.05	11.21***
	(2.88)	(3.34)
Zpc11	3.99	1.60
	(2.91)	(3.39)
Zpc12	0.46	2.78
	(2.99)	(3.45)
Zpc13	−0.55	14.34***
	(2.89)	(3.38)
Zpc14	0.95	5.37
	(2.83)	(3.35)
Zpc15	3.52	9.34**
	(2.82)	(3.34)
R^2	0.11	0.28
Adj. R^2	0.10	0.27
Num. obs.	1202	619
RMSE	4.89	5.70

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4.1: **Regression of 15 principal component vectors to DM on White and Yellow cassava.**

white and yellow respectively; 15 PCs explained 10% and 27% of the total DM variation in white and yellow respectively (Table 4.1) while 50 PCs explained 20% and 31% of the total DM variation in white and yellow respectively. Ideally, significant PCs from these analyses should be the PC vectors used in the RcHM analysis.

4.5.3 Sensitivity analysis:

The observed differences between segment correlations for the RcHM analysis with -0.5, 0 and 0.5 genomic correlations (Figures 5.6, 5.7 and 5.8) show their sensitivity to changes in these genomic correlations. These changes in segment correlations for the three subpopulations reflect sensitivities of segment correlations to changes in whole genome correlations.

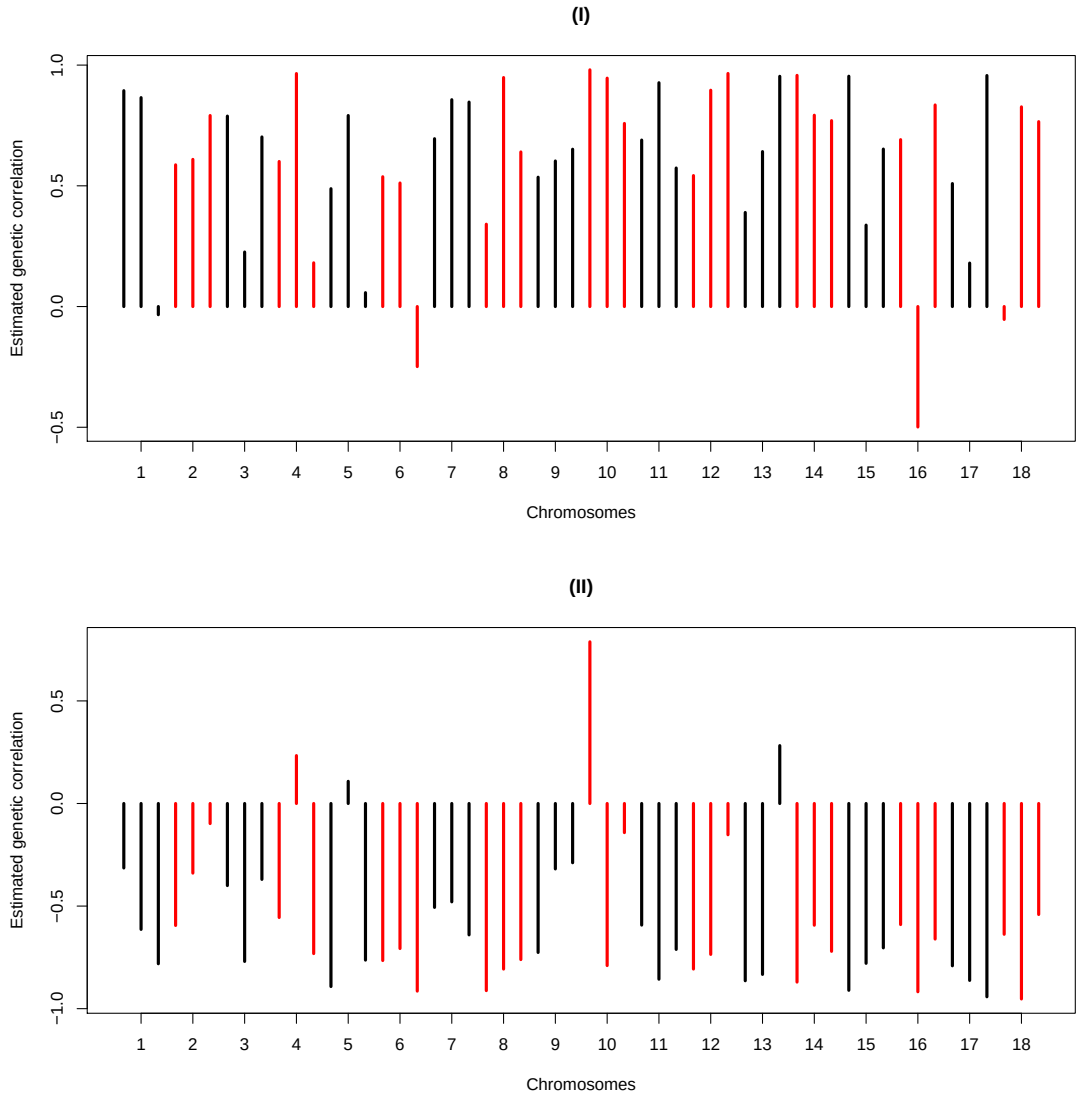

Figure 4.6: **Sensitivity of segment correlations at fixed genome-wide genetic correlation values of 0.5 and -0.5 for white cassava.** Top and bottom plots show segment correlations (bars) between DM and B for fixed genomic correlations 0.5 and -0.5 respectively.

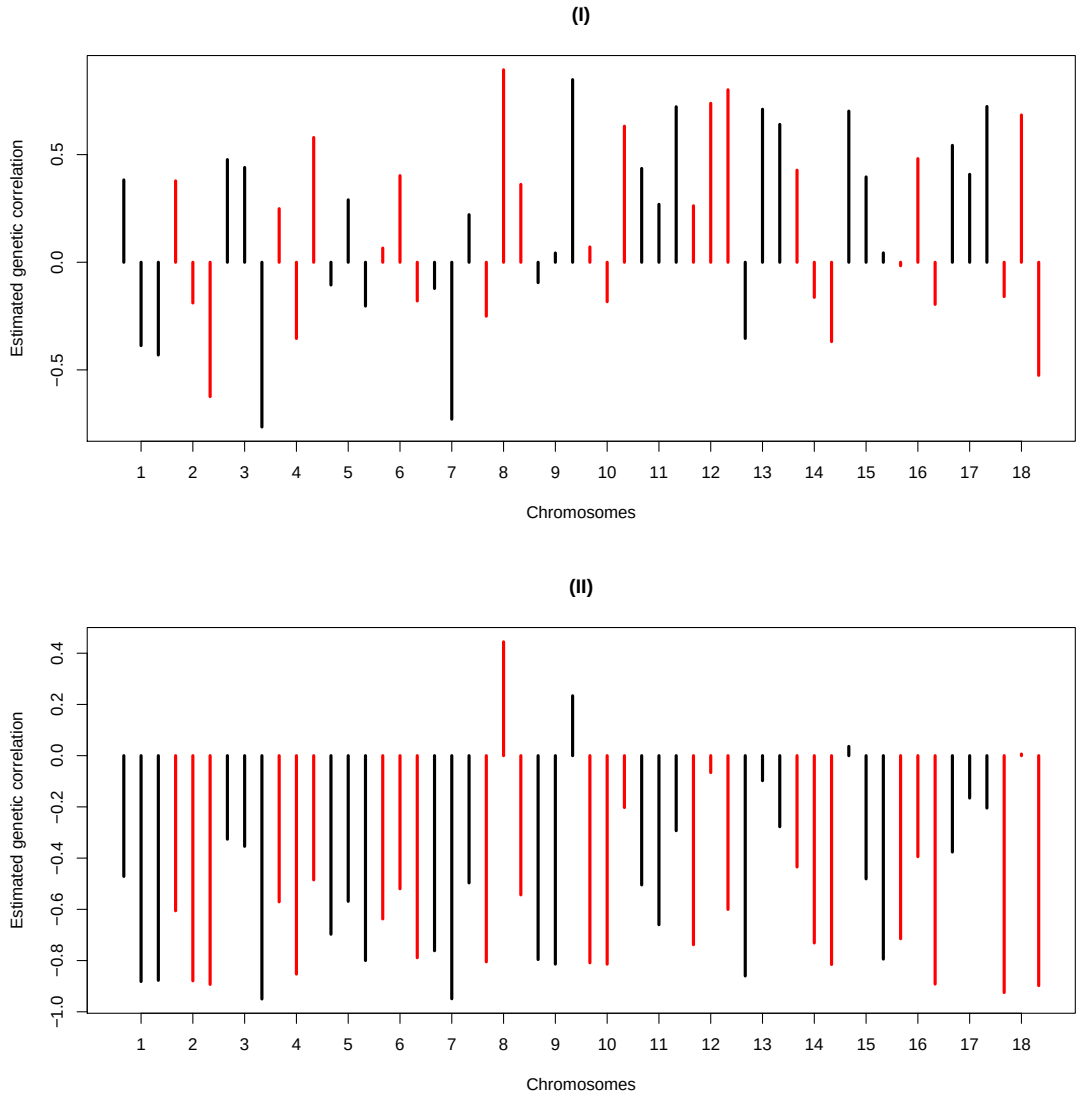

Figure 4.7: **Sensitivity of segment correlations at fixed genome-wide genetic correlation values of 0.5 and -0.5 for yellow cassava.** Top and bottom plots show segment correlations (bars) between DM and B for fixed genomic correlations 0.5 and -0.5 respectively.

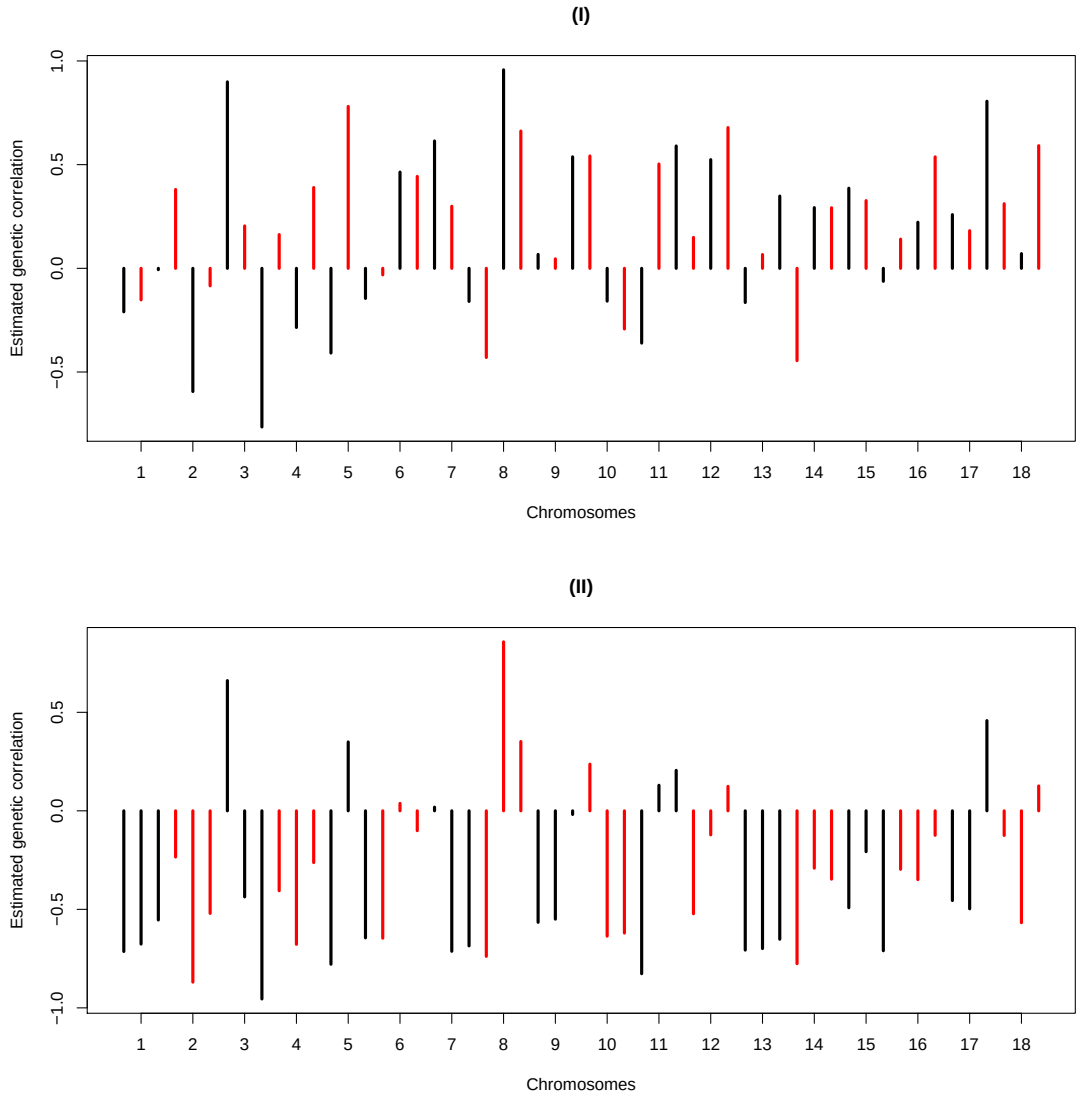

Figure 4.8: **Sensitivity of segment correlations at fixed genome-wide genetic correlation values of 0.5 and -0.5 for yellow-plus-white cassava.** Top and bottom plots show segment correlations (bars) between DM and B for fixed genomic correlations 0.5 and -0.5 respectively.

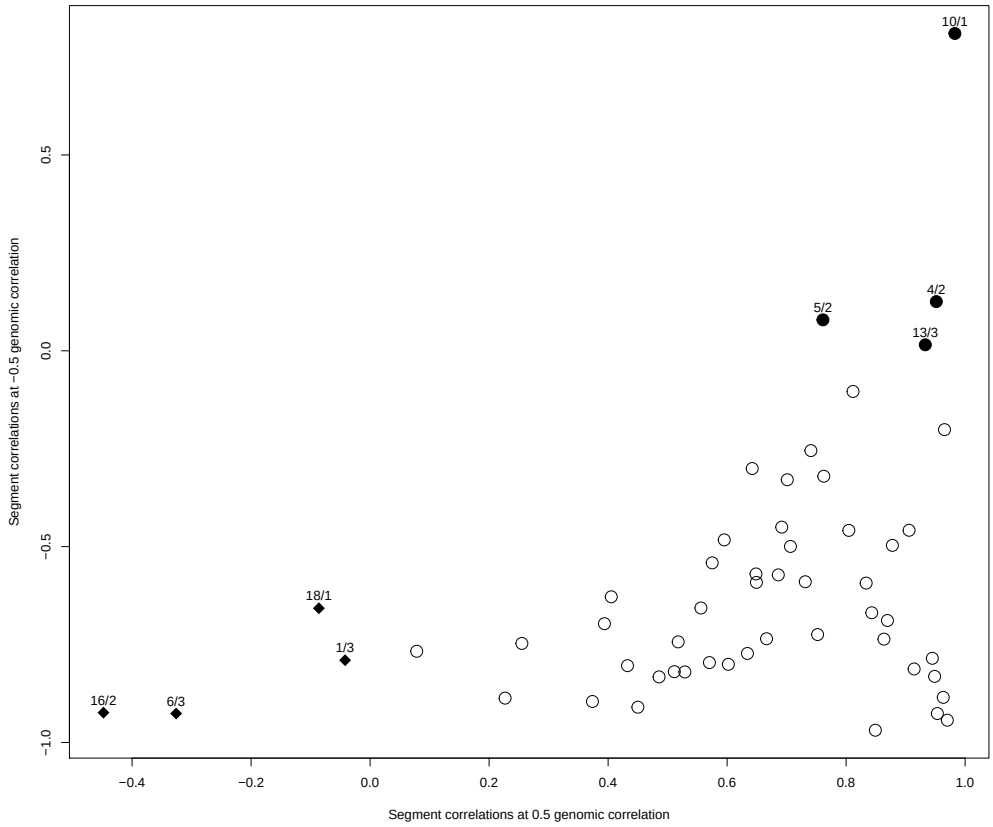

Figure 4.9: **Genomic segment correlations between 0.5 and -0.5 genome-wide genetic correlations in white cassava.** Scatter plot shows consistencies and differences between segment correlations when fixed genome-wide genetic correlations were changed from -0.5 to 0.5 for DM and B in white cassava. Bold circles represent consistent favorable (positive) segments and bold diamonds for consistent unfavorable segments. Locations of some segments are also shown for example 10/1 is the first segment on chromosome 10.

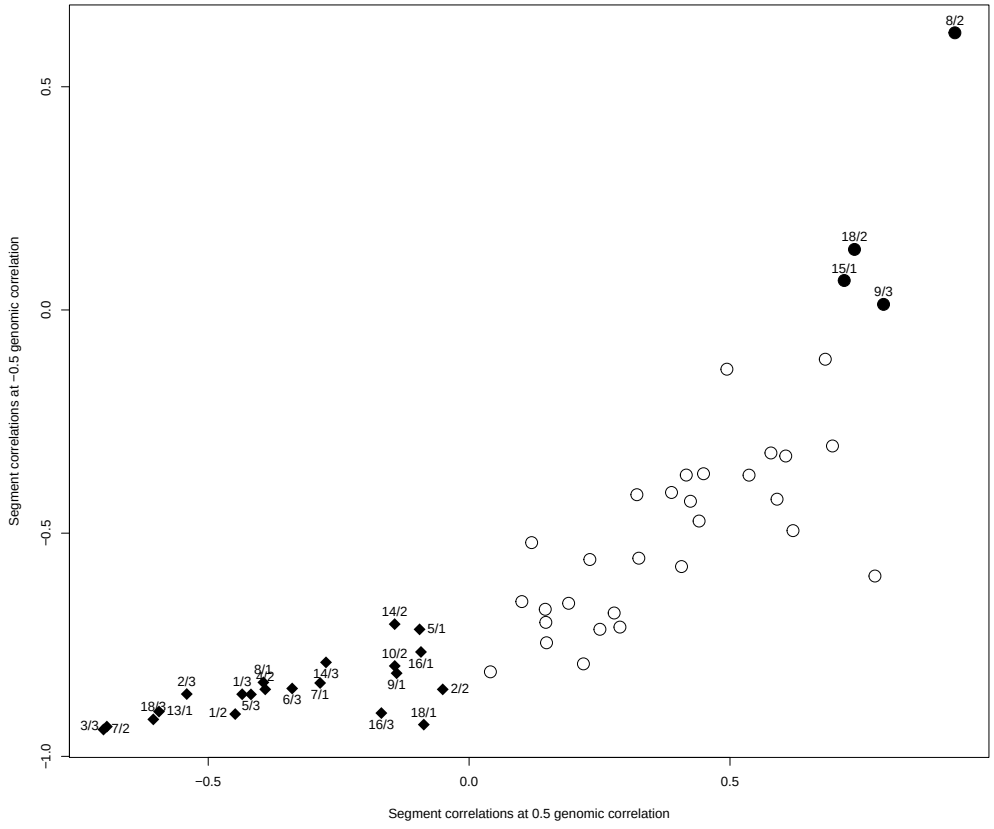

Figure 4.10: **Genomic segment correlations between 0.5 and -0.5 genome-wide genetic correlations in yellow cassava.** Scatter plot is as described in Figure 4.9 but for yellow cassava.

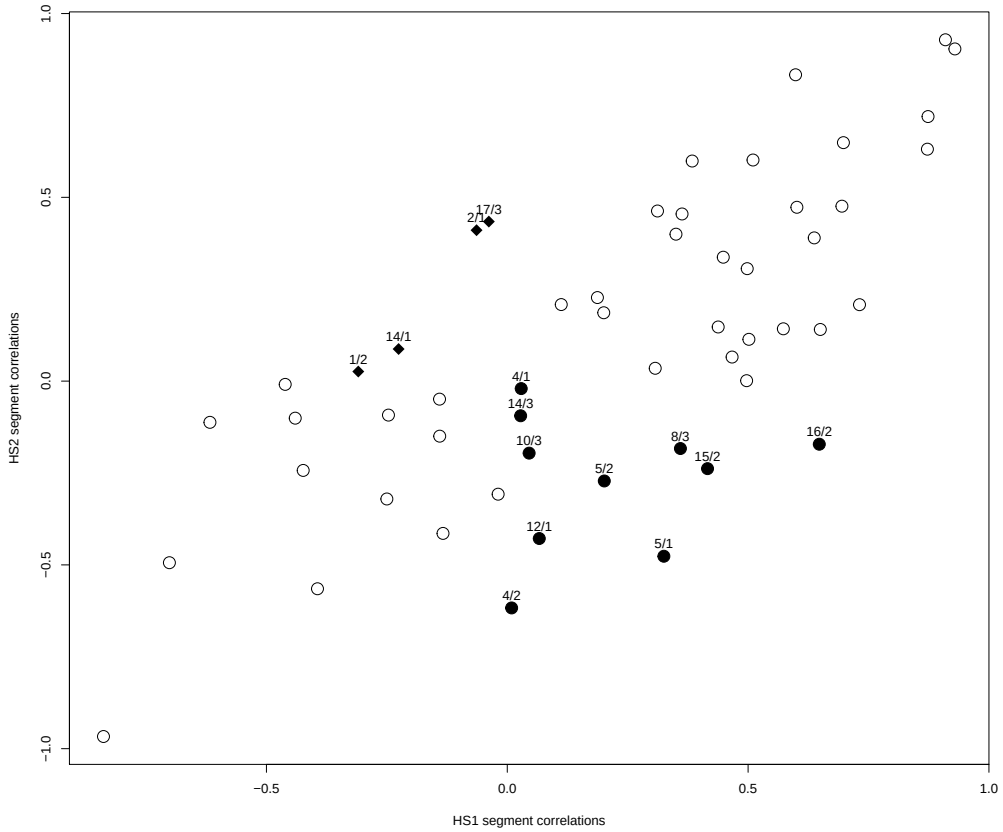

Figure 4.11: **Relationship between segment correlations and time of harvest for white cassava.** Scatter plot shows consistencies and differences between segment correlations when harvest time changes from 12 MAP to 14 MAP for DM and FYLD in white high starch cassava. Bold circles represent favorable segments in HS1 and bold diamonds for favorable segments in HS2. Segment locations were as described in Figure 4.9.

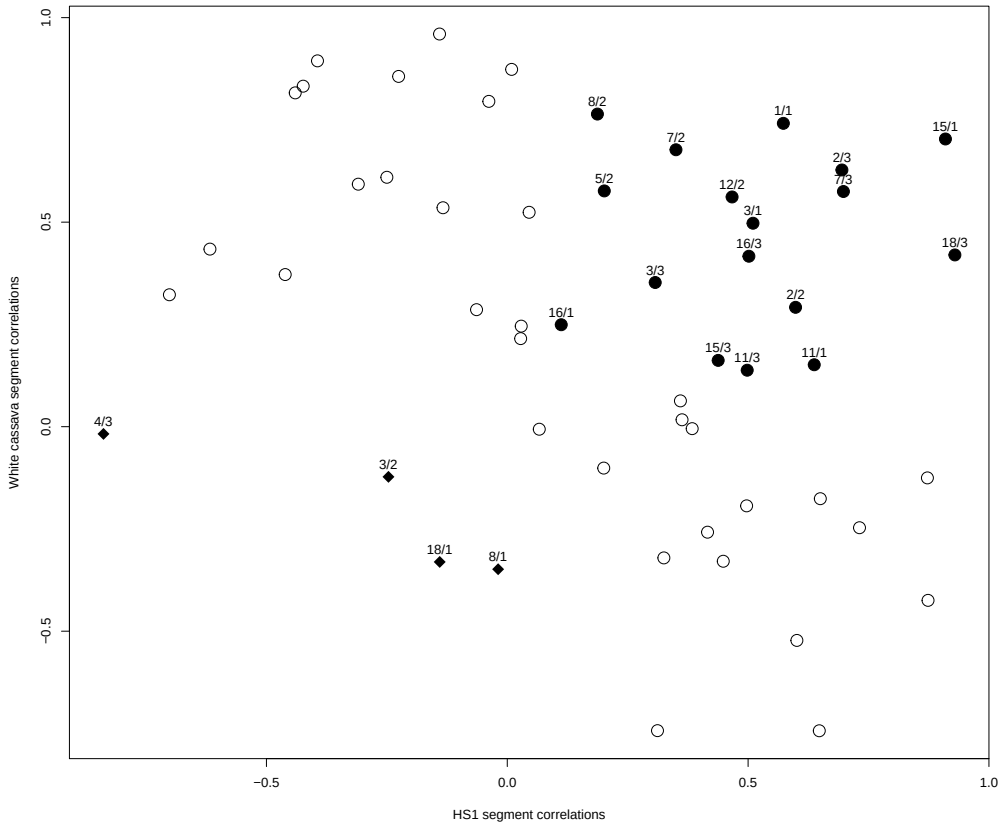

Figure 4.12: **Relationship between DM, B and FYLD in white cassava.** Scatter plot shows consistencies and differences between segment correlations for DM, B and FYLD in white cassava. Bold circles represent consistent favorable (positive) segments and bold diamonds for consistent unfavorable segments. Segment locations were as described in Figure 4.9.

We also observed that some segments were consistently favourable or unfavorable in the white germplasm even when genomic correlations between DM and B changed from -0.5 to 0.5 (Figure 4A). These include favourable segments on Chromosomes 4, 5 and 10 and unfavourable segments in chromosomes 1, 6, 16 and 18. Significant segments associated with DM on white cassava were previously identified on chromosomes 1, 4, 5, 10 and 18 by Okeke et al., 2017. These further reflect the impact of genomic segments on these chromosomes for DM accumulation in white cassava germplasm. However for yellow cassava, we observed that 39% of genomic segments in yellow were consistently negative (unfavourable) even when genomic correlations between DM and B changed from -0.5 to 0.5 (Figure 4B). These include segments from all chromosomes in the cassava genome excluding chromosomes 11, 12 and 17. This further reflects the complexity of improving DM in yellow cassava germplasm.

The RcHM procedure described herein is a multi-stage procedure but may be beneficial for mapping co-inheritance in complex traits as we have showed for some cassava traits. However, the ideal bivariate mixed model for the RcHM would be a multikernel model with the kernels as genomic relationship matrices calculated from (a) SNPs in the target segment and (b) from remaining whole genome SNPs. We tried this model but without success mainly due to convergence issues which may have arisen from our limited data.

4.6 Conclusion:

We have shown herein that the RcHM is a powerful approach that can be used to identify genomic segments that strongly affect the co-inheritance of two complex traits. The RcHM procedure produces a genome-wide correlation map of genomic segments. These segments represent large blocks of the genome that may be inherited as a unit (haplotypes) and which harbor QTL variants implicated in complex traits. Bivariate SNP effects were used to obtain GSVs which may provide better estimates of inheritance units. Correlation estimates of GSVs from two traits harbour combined associations of segments to both traits revealing the magnitude and direction of the segments association to both traits. If this correlation is favourable, then the segment might be of interest to the breeder and CSVs may be used in a multitrait index for selection. However, an interesting flip side is that genomic segment correlation maps can give insights into the co-inheritance profile of two traits and as a result provide a better understanding of how the traits should be improved. In this study, genomic segment correlation maps showed a favourable path for the development of high DM white cassava varieties while a limited prospect was shown for development of high DM yellow cassava. We advocate for more research into the use of CSVs in multitrait merit indices and conclude that RcHM is beneficial for understanding the co-inheritance of complex traits.

Table 4.2: **Genetic parameters for different populations of African cassava.**

	Genetic variance				Heritability	
Population	**DM**	**B**	**Genetic correlation (DM, B)**	**Error correlation (DM, B)**	**DM**	**B**
White	11.72 (1.62)	0.58 (0.11)	0.14 (0.010)	-0.07 (0.005)	0.57	0.11
Yellow	12.55 (3.88)	6.46 (1.45)	-0.30 (0.012)	-0.05 (0.009)	0.44	0.47
Yellow-plus-white	15.77 (0.89)	7.99 (0.31)	-0.08 (0.001)	-0.05 (0.001)	0.62	0.57
	DM	**FYLD**	**Genetic correlation (DM, FYLD)**	**Error correlation (DM,FYLD)**	**DM**	**FYLD**
White (HS1)	8.79 (3.02)	9.21 (3.68)	0.09 (0.02)	0.08 (0.03)	0.32	0.12
White (HS2)	7.16 (2.53)	11.39 (4.49)	0.12 (0.03)	-0.007 (0.02)	0.26	0.16

Traits are dry matter (DM), root yellowness (B), and fresh root yield (FYLD). Standard errors for parameter estimates are in parenthesis.

4.7 References:

Aguilar, I., Misztal, I., Tsuruta, S., Legarra, A. and Wang, H., 2014. PREGSF90POSTGSF90: computational tools for the implementation of single-step genomic selection and genome-wide association with ungenotyped individuals in BLUPF90 programs. Proc. 10th World Congr. Genet. Appl. Livest. Prod.

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

Ayres, D.L., Darling, A., Zwickl, D.J., Beerli, P., Holder, M.T., Lewis, P.O., Huelsenbeck, J.P., Ronquist, F., Swofford, D.L., Cummings, M.P. and Rambaut, A., 2011. BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. Systematic biology, p.syr100.

Bauer, A.M. and Lon, J., 2008. Multiple-trait breeding values for parental selection in self-pollinating crops. Theoretical and Applied Genetics, 116(2), pp.235-242.

Bouis, H., Biofortification Progress Briefs August 2014. Washington DC.: Harvest Plus; 2014 August 2014. 82 p.

Bouis, H.E. and Welch, R.M., 2010. Biofortificationa sustainable agricultural strategy for reducing micronutrient malnutrition in the global south. Crop Science, 50(Supplement_1), pp.S-20.

Bouis, H.E., Hotz, C., McClafferty, B., Meenakshi, J.V. and Pfeiffer, W.H., 2011. Biofortification: a new tool to reduce micronutrient malnutrition. Food and nutrition bulletin, 32(1_suppl1), pp.S31-S40.

Broadbent, A.D., 2004. A critical review of the development of the CIE1931 RGB colormatching functions. Color Research and Application, 29(4), pp.267-272.

Carter, A.J. and Nguyen, A.Q., 2011. Antagonistic pleiotropy as a widespread mechanism for the maintenance of polymorphic disease alleles. BMC medical genetics, 12(1), p.160.

Ceballos, H., Morante, N., Snchez, T., Ortiz, D., Aragon, I., Chvez, A.L., Pizarro, M., Calle, F. and Dufour, D., 2013. Rapid cycling recurrent selection for increased carotenoids content in cassava roots. Crop Science, 53(6), pp.2342-2351.

Cole, J.B. and VanRaden, P.M., 2011. Use of haplotypes to estimate Mendelian sampling effects and selection limits. Journal of Animal Breeding and Genetics, 128(6), pp.446-455.

Daetwyler, H.D., Hayden, M.J., Spangenberg, G.C. and Hayes, B.J., 2015. Selection on optimal haploid value increases genetic gain and preserves more genetic diversity relative to genomic selection. Genetics, 200(4), pp.1341-1348.

Douc, R. and Capp, O., 2005, September. Comparison of resampling schemes for particle filtering. In Image and Signal Processing and Analysis, 2005. ISPA 2005. Proceedings of the 4th International Symposium on (pp. 64-69). IEEE.

Ebah-Djedji, B.C., Dje, K.M., N'Zue, B., Zohouri, G.P. and Amani, N.G., 2012. Effect of harvest period on starch yield and dry matter content from the rootous roots of improved cassava (Manihot esculenta Crantz) varieties. Pak. J. Nutr, 11(5), pp.414-418.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S. and Mitchell, S.E., 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS one, 6(5), p.e19379.

Esch, E., Szymaniak, J.M., Yates, H., Pawlowski, W.P. and Buckler, E.S., 2007. Using crossover breakpoints in recombinant inbred lines to identify quantitative trait loci controlling the global recombination frequency. Genetics, 177(3), pp.1851-1858.

Esuma, W., Kawuki, R.S., Herselman, L. and Labuschagne, M.T., 2016. Diallel analysis of provitamin A carotenoid and dry matter content in cassava (Manihot esculenta Crantz). Breeding science, 66(4), pp.627-635.

Ferreira, M.A. and Purcell, S.M., 2009. A multivariate test of association. Bioinformatics, 25(1), pp.132-133.

Frszczak, M. and Szyda, J., 2016. Comparison of significant single nucleotide polymorphisms selections in GWAS for complex traits. Journal of applied genetics, 57(2), pp.207-213.

Galesloot, T.E., Van Steen, K., Kiemeney, L.A., Janss, L.L. and Vermeulen, S.H., 2014. A comparison of multivariate genome-wide association methods. PloS one, 9(4), p.e95923.

Glaubitz, J., T. Casstevens, R. Elshire, J. Harriman, and E.S. Buckler. 2012. TASSEL 3.0 genotyping by sequencing (GBS) pipeline documentation. Edward S. Buckler, USDA-ARS, Ithaca, NY.

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., and Rokhsar, D. S. (2012). Phytozome: a comparative platform for green plant genomics. Nucleic acids research, 40(D1), D1178-D1186. (http://www.phytozome.net; accessed 1 July, 2014).

Hammer, G.L., Hobman, F.R. and Shepherd, R.K., 1987. Effects of planting time and harvest age on cassava (Manihot esculenta) in Northern Australia. I. Crop growth and yield in moist environments. Experimental agriculture, 23(04), pp.401-414.

Hardle, W. and Marron, J.S., 1991. Bootstrap simultaneous error bars for non-parametric regression. The Annals of Statistics, pp.778-796.

Henderson, C.R., 1973. Sire evaluation and genetic trends. Journal of Animal Science, 1973(Symposium), pp.10-41.

Iglesias, C., Mayer, J., Chavez, L. and Calle, F., 1997. Genetic potential and stability of carotene content in cassava roots. Euphytica, 94(3), pp.367-373.

Kawano, K., Fukuda, W.M.G. and Cenpukdee, U., 1987. Genetic and environmental effects on dry matter content of cassava root. Crop Science, 27(1), pp.69-74.

Kemper, K.E., Bowman, P.J., Pryce, J.E., Hayes, B.J. and Goddard, M.E., 2012. Long-term selection strategies for complex traits using high-density genetic markers. Journal of dairy science, 95(8), pp.4646-4656.

Kemper, K.E., Saxton, S.J., Bolormaa, S., Hayes, B.J. and Goddard, M.E., 2014. Selection for complex traits leaves little or no classic signatures of selection. BMC genomics, 15(1), p.246.

Koivula, M., Strandn, I., Su, G. and Mntysaari, E.A., 2012. Different methods to calculate genomic predictionsComparisons of BLUP at the single nucleotide polymorphism level (SNP-BLUP), BLUP at the individual level (G-BLUP), and the one-step approach (H-BLUP). Journal of dairy science, 95(7), pp.4065-4073.

Leon, K., Mery, D., Pedreschi, F. and Leon, J., 2006. Color measurement in L a b units from RGB digital images. Food research international, 39(10), pp.1084-1091.

Luby, J.J. and Shaw, D.V., 2009. Plant breeders' perspectives on improving yield and quality traits in horticultural food crops. HortScience, 44(1), pp.20-22.

Ly, Delphine, et al. Relatedness and genotype x environment interaction affect prediction accuracies in genomic selection: A study in cassava. Crop Science53.4 (2013): 1312-1325.

Ma, X., Zhu, Q., Chen, Y. and Liu, Y.G., 2016. CRISPR/Cas9 platforms for genome editing in plants: developments and applications. Molecular plant, 9(7), pp.961-974.

Masuda, Y., Aguilar, I., Tsuruta, S. and Misztal, I., 2015. Technical note: Acceleration of sparse operations for average-information REML analyses with supernodal methods and sparse-storage refinements. Journal of animal science, 93(10), pp.4670-4674.

Maziya-Dixon, B., A.G.O. Dixon, and A.-R.A. Adebowale. 2007. Targeting different end uses of cassava: Genotypic variations for cyanogenic potentials and pasting properties. Int. J. Food Sci. Technol. 42:969976. doi:10.1111/j.1365-2621.2006.01319.x

Maziya-Dixon, B., Kling, J.G., Menkir, A. and Dixon, A., 2000. Genetic variation in total carotene, iron, and zinc contents of maize and cassava genotypes. Food and Nutrition Bulletin, 21(4), pp.419-422.

Okechukwu, R.U., and A.G.O. Dixon. 2008. Genetic gains from 30 years of cassava breeding in Nigeria for storage root yield and dis-

ease resistance in elite cassava genotypes. J. Crop Improv. 22:181208. doi:10.1080/15427520802212506

Okeke, G.U., Akdemir, D., Rabbi, I., Kulakow, P. and Jannink, J. 2017. Regional Heritability Mapping provides insights into Dry matter (DM) Content in African white and yellow cassava populations. The Plant Genome (in press).

Oparinde, A., Banerji, A., Birol, E. and Ilona, P., 2016. Information and consumer willingness to pay for biofortified yellow cassava: evidence from experimental auctions in Nigeria. Agricultural Economics.

Sayre, R., Beeching, J.R., Cahoon, E.B., Egesi, C., Fauquet, C., Fellman, J., Fregene, M., Gruissem, W., Mallowa, S., Manary, M. and Maziya-Dixon, B., 2011. The BioCassava plus program: biofortification of cassava for sub-Saharan Africa. Annual review of plant biology, 62, pp.251-272.

Strandn, I. and Garrick, D.J., 2009. Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. Journal of dairy science, 92(6), pp.2971-2975.

Tanumihardjo, S.A., Bouis, H., Hotz, C., Meenakshi, J.V. and McClafferty, B., 2008. Biofortification of staple crops: an emerging strategy to combat hidden hunger. Comp Rev Food Sci Food Safety, 7, pp.329-34.

Team, R.C., 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013.

VanRaden, P.M., 2008. Efficient methods to compute genomic predictions. Journal of dairy science, 91(11), pp.4414-4423.

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genetics 2: e190.

CHAPTER 5

CONCLUSION:

The overall aim of this study was towards the genetic improvement of African cassava using GS. GS promises to accelerate gains for crops like cassava and this is critical for food security in areas of the world where cassava is a staple. Conclusions from this study is structured as recommendations below:

5.1 Cassava Genetic evaluation:

We advocate that the models used in this work (Chapter 1) be evaluated using predicted selection gains in addition to the usual prediction accuracies used in literature. The genetic correlations from the ME model can help the breeder to understand the expected correlated responses to selection when clones selected at the central breeding station Ibadan are deployed in other cassava production regions. The uE model in this study was equivalent to a model with compound symmetry covariance structure which leads to a more parsimonious approach than the ME model. However, the differences between the ME and uE models will be seen more clearly when predicted selection gains are estimated. We recommend multivariate models for genetic evaluation in cassava. A model that combines the benefit of the MT and ME models may be beneficial.

We further believe that GS is very useful for cassava breeding especially at the critical stage of CETs. In the NextGen project, clones from same families were evaluated at different locations as a form of replication. This was a very good approach and will help in capturing GxE as early as in CETs. Captur-

ing GxE in this stage is facilitated by the GRM which helps to better connect phenotypic data from families. This means that with uE and ME models, more accurate parental selections could be made because GxE information has been accounted for.

During the course of preparing this thesis, the author understood that at the early stages of setting up the cassava breeding program at the IITA, the emphasis were on disease resistance for CMD and CBB [1-3]. Consequently, a resistant clone from the successful interspecific hybridization of cassava and *Manihot glaziovii* was utilized heavily and influenced the development of what is now the Genetic gain population. Several discussions has led to thinking that genomic segments that conferred resistance to these diseases may also be placing some constraints on other traits that are now of interest to the breeding program. Cassava genetic evaluation systems should account for this and further investigations using independent germplasm like the local landraces should be a valuable resource to understand the impact of these segments.

5.2 Lessons from the mapping of complex traits:

In this study, we found that the RHM and RcHM were powerful approaches for understanding the genetic basis for inheritance or co-inheritance respectively for complex traits in cassava. The RcHM showed that concurrent improvement of DM and FYLD can be made in white cassava using an index based on multivariate breeding values. However a limited prospect was shown for improving DM content in yellow cassava. This presents both a challenge and an opportunity to re-organize the yellow cassava breeding program. We advocate that

as early as in the seedling nursery, an index based on the EGVs from an MT model should be used for selection for DM and tuber yellowness using continuous phenotypes from the near infra-red spectroscopy (NIR). This means that NIR calibrations for DM and beta-carotene should be pursued at the seedling nursery stage. This will complement hybridization with exotic germplasm from CIAT at a location with low CMD influence.

Genomic segments can be deployed in breeding using optimization procedures [4]. However, this needs further investigation.

5.3 Hybrid cassava breeding:

I advocate for a pilot scale program geared towards development of cassava hybrids. These imply development of selfing procedures and heterotic groups. Again, the selection criteria for selfed lines should be based on an index from multivariate breeding values. I believe that this will facilitate the development of different product profiles including varieties for different types of cassava starch which is now in high demand in the international market.

5.4 Lessons for fellow young scientists:

This study has deeply exposed the author to the world of quantitative genetics and genetic evaluation. The power and efficiency of linear mixed models for gleaning information from phenotypic and genotypic data can not be overemphasized. For young scientists seeking to delve into the field of quantitative

genetics, a very good understanding of matrix algebra, mathematical statistics and technical computing is absolutely required. The expertise on data analysis is one that will continue to be in much demand especially in this era of big data in agriculture. I will continue the work on the development of EMMREML [5] and also developing models using the BLUPf90 family of programs [6].

5.5 References:

[1] Hahn, S.K., Howland, A.K. and Terry, E.R., 1973. Cassava Breeding at IITA. IITA,[sd].

[2] Hahn, S.K. and Howland, A.K., 1972. Breeding for resistance to cassava mosaic. In Proceedings IDRC/IITA Cassava Mosaic Workshop, International Institute of Tropical Agriculture. Ibadan, Nigeria.

[3] Hahn, S.K., Terry, E.R. and Leuschner, K., 1980. Breeding cassava for resistance to cassava mosaic disease. Euphytica, 29(3), pp.673-683.

[4] Goiffon, M., Kusmec, A., Wang, L., Hu, G. and Schnable, P., 2017. Improving Response in Genomic Selection with a Population-Based Selection Strategy: Optimal Population Value Selection. Genetics, pp.genetics-116.

[5] Akdemir, D. and Okeke, U.G., 2015. EMMREML: Fitting mixed models with known covariance structures. R package version 3.1.

[6] Misztal, I., Tsuruta, S., Strabel, T., Auvray, B. and Druet, T., 2007. BLUPF90 family of programs. University of Georgia. http://nce. ads. uga. edu/ ignacy/numpub/blu pf90/, Accessed on Jan, 2, p.2007.