HIGH-THROUGHPUT PHENOTYPING AND GENOMICS-ASSISTED BREEDING FOR QUALITY TRAITS IN 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

Ugochukwu Nathaniel Ikeogu

August 2018

HIGH-THROUGHPUT PHENOTYPING AND GENOMICS-ASSISTED BREEDING FOR QUALITY TRAITS IN CASSAVA

Ugochukwu Nathaniel Ikeogu, Ph. D.

Cornell University 2018

To promote rapid and standardized phenotyping for genomic improvement of quality traits in cassava, calibrations for dry matter content (DMC) and carotenoids in fresh cassava roots were developed from a portable near infra-red spectrometer (NIRS). Effect of eight pre-treatment combinations was evaluated on calibration performance and standard normal variate and de-trend (SNVD), with the first derivative calculated on two data points and no smoothing (SNVD+1111), was adequate to build a robust model. Generally, high calibration performance was obtained for most traits e.g. model for DMC on mashed samples had - R2c = 99%, R2cv = 95%, RPD = 4.5 and SECV = 0.9, with satisfactory R2 of 80% on independent validation set. On average, models developed with mashed were better than the intact samples. Intact and mashed NIRS-derived DMC were highly correlated (0.94) and had higher correlations (>0.95) with the ideal oven-drying than the specific gravity methods (0.49 and 0.69, depending on the dataset). Non-linear calibration model using random forest (RF), was equally develop and used to process spectra from National Root Crops Research Institute (NRCRI), Umudike for carotenoids including total carotenoid content (TCC) and some individual carotenoids (ICS): all-trans β-carotene (ATBC), violaxanthin (VIO), Lutein (LUT), 15-Cis beta-carotene (15CBC), 13-Cis beta-carotene (13CBC), Alpha-carotene (AC), 9-Cis beta-carotene (9CBC) and phytoene (PHY) . Derived carotenoids were used to understand correlations (phenotypic and genotypic), especially between TCC and ICS. High and positive phenotypic and genotypic correlations (>0.75) were obtained between TCC and the ICS except for PHY and LUT. Genome-wide association studies identified previously reported region on chromosome 1 associated with variation in TCC, in addition to other unidentified associations for both TCC and the ICS.

Evaluating the potential of using Genome-wide predictions for carotenoids improvement, higher predictions were obtained from non-linear RF model with a one-step approach in single and multi-trait scenarios than linear and two-step approaches. The possibility of using molecular markers to assign parentage to progenies from a polycross nursery scheme was demonstrated with 100% assignment accuracy from simulated datasets. The information provided in this study is vital in redefining cassava breeding.

BIOGRAPHICAL SKETCH

Ugochukwu Ikeogu was born in Umuasua, Isuikwuato LGA of Abia State in South East Nigeria on October 15, 1982. He had his Primary and Secondary education at Amaba Central School, Amaba and Secondary Technical School, Ovim, respectively, both in Isuikwuato LGA. He obtained both his Bachelor in Agronomy and Masters in Plant Breeding and Genetics from the Department of Agronomy, College of Crop and Soil Sciences, Michael Okpara University of Agriculture, Umudike (MOUAU), Nigeria. His interest in Breeding and Genetics intensified after visiting a *Musa* Breeding station at IITA Onne, Rivers State, Nigeria. He was electrified seeing Banana/Plantain seeds for the first time and how breeding could be used in developing new food products. He grew up among Professional and committed Agriculturists and from childhood, had enjoyed having his personal vegetable garden and taking care of livestock and poultry birds. In 2010, he took up a Research Assistant position at the National Root Crops Research Institute, Umudike under the Generation Challenge Project (GCP), where he was responsible for designing and managing experimental trials, data collection and analyses as well as coordinating hybridization activities. He later joined the Next Generation Project in 2013 and got admitted into Cornell University the same year for his Ph.D. In the course of his study, he was involved in a lot of collaborations across many institutions including CIAT, IITA etc., and was passionate in developing tools for efficient cassava breeding.

He sees himself as a product of God's grace and having received mercy, he has passion in helping the less privileged, especially children to acquire quality education. He is delighted in translating theoretical knowledge to tangible products for food security and economic empowerment.

Dedicated to Bill and Melinda Gates for their benevolence and dedication in improving the quality of life of individuals around the globe.

ACKNOWLEDGMENTS

I am indebted and grateful to Next generation (NEXTGEN) cassava initiators and funding agencies, the Bill & Melinda Gates Foundation and the UK Department for International Development, for the opportunity to pursue my Ph.D.

I express my deepest gratitude to the members of my committee, Prof. Jean-Luc Jannink, Prof. Susan McCouch and Prof. Donald Viands for their endless support and mentorship during the course of my study.

I appreciate the efforts of the past and current leadership of NextGen including the External Project Advisory Committee – Prof. Ronnie Coffman, Dr. Hale Tufan, Prof. Chiedozie Egesi, Dr. Ben Hayes, Dr. Steve Rounsley, Dr. David Meyer, Prof. Edward Buckler, etc. for their invaluable contribution towards my work.

Special thanks to all the collaborating institutions - The International Center for Tropical Agriculture (CIAT), Cali, Colombia; International Institute of Tropical Agriculture (IITA,) Ibadan, Nigeria; Agricultural Research for Development (CIRAD), France and Colombia; Boyce Thompson Institute, Ithaca, USA and National Root Crops Research Institute (NRCRI), Umudike, Nigeria including Dr. Hernan Ceballos, Dr. Peter Kulakow, Dr. Ismail Rabbi, Dr. Dominique Dufour, Dr. Fabrice Davrieux, Dr. Lukas Mueller, Dr. Joseph Onyeka, John Belalcazar, Ahamefula Nwogu, Princess Onyegbule, Alex Ogbonna among many others, for creating an enabling environment to do my work.

I value the contributions of the past and present members of Sorrells and Jannink's lab especially, Prof. Mark Sorrells, Dr. Deniz Akdemir, Dr. Uche Okeke, Dr. Marnin Wolfe,

Dr. Roberto Lozano, Dr. Mohammed Ibrahim etc., as well as the Plant Breeding and Genetics section for creating a positive academic and research environment.

I deeply appreciate all my family members: my wife - Dr. Nnenna Ikeogu, my parents – Dr. John Ikeorgu and Pastor (Mrs) Etochi Ikeorgu, siblings, in-laws and many other relatives for their encouragement and emotional support in the course of my study.

And unto God, the source of my strength and hope, be all glory, power, wisdom and dominion for His faithfulness. His mercy kept me.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

Chapter 2 has been published as an original research article in PloS ONE:

Ikeogu UN, Davrieux F, Dufour D, Ceballos H, Egesi CN, Jannink J-L (2017). Rapid analyses of dry matter content and carotenoids in fresh cassava roots using a portable visible and near infrared spectrometer (Vis/NIRS). PLoS ONE 12(12): e0188918. https://doi.org/10.1371/journal.pone.0188918

Chapter 3 will be submitted as an original research article to (undecided):

Ikeogu UN, et al. (Ready for submission). Non-Linear Calibration of Portable and Near Infrared Spectrometer (Vis/NIRS) in Trait Correlations, Genome-Wide Association Studies and Genomic Predictions of Carotenoids concentration in Cassava Roots.

Chapter 4 will be submitted as an original research article to (undecided):

Ikeogu UN, et al. (Ready for submission). SNP-Based Parentage Evaluation of a Polycross Nursery in the Genomic Selection Breeding Scheme of Cassava.

CHAPTER 1: GENERAL INTRODUCTION

Cassava is an important component and a reliable source of calories in the diets of more than 800 million people around the world (FAO 2013; Halsey et al. 2008). It is an outstanding food security crop because of its distinct morphological and physiological attributes (Barratt et al. 2006; Iglesias et al. 1997). Cassava grows well under marginal conditions and tolerates acidic soils where few other crops could hardly survive. It is generally a drought tolerant crop with the ability to thrive in low fertility and low P-conditions. It is also known to be resistant to most important diseases and pests (El-Sharkawy 2012). More importantly, cassava offers the most convenient harvest flexibility to farmers such that it can be kept in the field until farmers are ready to harvest it (Hernán Ceballos et al. 2004).

Cassava belongs to the family Euphorbiaceae and is an annual crop that is mostly propagated by stem (Nair et al. 2015; Chibueze Izah, Bassey, and Ohimain 2018). Depending on varieties and farmers' need, harvesting can be between 7 – 13 months after planting and can be reserved in the soil for up 2 - 3 years of planting (Chibueze Izah, Bassey, and Ohimain 2018). Cassava as a rich source of starch is generally utilized as food and feed for animals. It has a great but underexplored potentials as raw materials for many industrial purposes. The drive to maximize these potentials is currently attracting a lot of attention in terms of research and investment. It has a potential role as a source of raw materials for the biofuel industry (Egesi et al. 2007; Ceballos et al. 2017).

Cassava production is mostly subsistence, predominated by peasant, low income farmers whose production output is considerably low and barely enough for food and income generation (Cock 1987; Kawano 2003). Recently, cassava production has been on the increase and will continue to increase as the demand for more cassava products both in quantity and quantity increases (Anyanwu et al. 2015). Production in Asia has been characterized by increase in yield per unit area, strongly driven by the demand for dried cassava and starch for livestock feed and industrial applications, whereas increase in land area per hectare is mainly associated with production in Africa, driven by expanding urban markets for food products (FAO 2013). Cassava has huge potentials for further production increase. Under optimal conditions in experimental stations, fresh root yields of up to 80 t/ha per year and up to 60 t/ha per year under farm conditions are feasible (Kawano, Fukuda, and Cenpukdee 1987). This is still far greater than less than 20 t/ha per year of the current world average yield (United Nations 2015; FAO 2013).

There is a booming global demand for cassava and cassava products (Jansson C, Westerbergh A, Zhang J, Hu X 2009), which offers millions of cassava growers in tropical countries the opportunity to intensify production, earn higher incomes and boost the food supply within and outside the sub-Saharan. Yield in cassava is usually realized in fresh weight of roots per given area or proportion of root weight over shoot weight referred to as harvest index. However, cassava roots are bulky with about 70% moisture content and require processing to extract the dry root matter which has been estimated to contain 70-90% starch and the remainder being fibers. Such raw or unmodified cassava starches are progressively significant in human food, textile, alcohol and animal industries world-wide (Henry, Westby, and Collinson 1998). Because of the bulky

nature of the crop, major emphasis on dry root matter over fresh root yield is highly desirable. This is because, an increase in dry matter content (equivalent to starch content) translates invariably into higher income per unit land and per unit labor (investment) for cassava growers (Fakir et al. 2012). Low yielding varieties are usually susceptible to major known biotic and abiotic stresses. Also, climate change creates an environment for new and evolving strains of plant pathogenic organisms.

Traditionally, genetic improvement of cassava has largely relied on information obtained from phenotypes and sometimes on pedigrees to estimate the breeding values of genetic materials (Ceballos et al. 2004). This approach has recorded a level of success although with some obvious challenges. The vegetative multiplication rate of cassava is low and usually from one plant and depending on the genotype, 5–10 cuttings typically can be obtained. This shortcoming implies a lengthy process to arrive at the point where replicated evaluations across several locations can be conducted. It takes about 5–6 years from the time the botanical seed is germinated until the evaluation/selection cycle reaches the regional trial stage when several locations can be included (Ojulong et al. 2008; Ceballos et al. 2004). Cassava breeding has equally relied greatly on mass phenotypic recurrent selection (Jennings and Iglesias 2002; Ceballos, Hershey, and Becerra-López-Lavalle 2012). Scanty information exists on general combining ability (GCA) effects (breeding value) for selection of parental materials. Little information is available at the early stages of selection and where it exists, there is usually no proper separation between GCA and SCA. Selection of parents has been substantially based on heterotic effects, which cannot be transferred sexually to the next generation. Under the current breeding schemes, large genetic loads are likely to remain hidden in cassava

populations and useful recessive traits are difficult to detect (Ceballos et al. 2004). Selection stages are usually based on non-replicated trials where a large proportion of genotypes are eliminated without the proper evaluation set up (Kawano 2003; Ceballos et al. 2004).

These foregoing arguments justify the need for further improvement on the efficiency and effectiveness of cassava breeding in order to meet the rising market demand. Advances in genomics, molecular biology, and statistical genetics have created a paradigm shift in crop genetic improvement techniques enabling new genomic-based strategies such as genome-wide association study (GWAS). GWAS detects causal genes or QTL from the association between genome-wide markers that are in linkage disequilibrium (LD) with the causal genes or QTL and phenotypes (Iwata et al. 2013; Spindel et al. 2015). GWAS has been useful in uncovering the genetic basis of some quantitative traits in cassava (Esuma et al. 2016; I. Y. Rabbi et al. 2017; Wolfe et al. 2016). Unlike GWAS, Genomic Selection (GS) is a form of marker-assisted selection in which genetic markers covering the whole genome are used so that all quantitative trait loci (QTL) are in linkage disequilibrium with at least one marker (Goddard and Hayes 2007). Genomic prediction combines marker data with phenotypic and pedigree data (when available) in an attempt to increase the accuracy of the prediction of breeding and genotypic values. This form of selection based on marker effects is gradually changing the traditional practices implored in animal breeding and is fast being adopted in plant breeding (Meuwissen, Hayes, and Goddard 2001; Heslot et al. 2012). Genomic prediction of breeding values involves a training analysis that predicts the influence of small genomic regions by a regression of observed information on marker genotypes

for a given population of individuals. Expectedly, large numbers of genotyped and phenotyped plants are required to produce robust estimates of the effects of SNPs that are summed together to generate genomic breeding values. This technique promises to speed up genetic gain leading to improved, higher yielding, broadly adapted, and stable genotypes. Despite these potentials, GS implementation and potential advantage remains to be realized in breeding programs especially where such resources are scarce. Besides GWAS and GS, the availability of abundant marker information is helping redefine important breeding techniques especially relating to mating and parentage assignment in breeding programs (Boerner 2017; Heaton et al. 2014; Van Eenennaam et al. 2007).

A major challenge in the adoption of new technologies in the 21st century is the need for quick and accurate quantification of many genetic materials. The effective development of GS models for the prediction of crop performance and the recurrent updating of such models require precise, standardized and low cost phenotyping tools for efficient dissection of quantitative traits. Near Infrared Spectroscopy (NIRS) is a technique that permits the screening of considerable large amounts of samples and the identification of qualitative and quantitative properties. The NIRS technology is built on the interaction of physical matter with the near infrared spectral region of light (Lopez et al. 2013). When compared to the conventional methods of phenotyping, NIRS is fast and easy to handle, versatile and can be used for the analysis of several traits simultaneously (Büning-pfaue 2001; Teye, Huang, and Afoakwa 2013). In addition, NIRS avoids high hazards and problems of organic and other chemical waste disposal.

NIRS is currently being used for the quantification of important traits in cassava (Sánchez et al. 2014; Lebot et al. 2009; Belalcazar, Dufour, Andersson, Pizarro, Luna, Londoño, Morante, Jaramillo, Pino, López-Lavalle, Davrieux, Talsma, and Ceballos 2016; Davrieux et al. 2016) and in many other fields (Marten, Shenk, and Barton 1989; Roggo et al. 2007; De Alencar Figueiredo et al. 2006; Xuan Zhang et al. 2013). It is becoming available in portable versions which provides more flexibility for field-based analyses.

This study evaluates the possibility of incorporating new phenotyping and genomic cutting edge tools to speed up genetic gains in cassava especially in a low-resource National Agricultural Research System (NARS) in Nigeria – National Root Crops Research Institute (NRCRI), Umudike. The country is strategic as the highest producer of cassava and growing population of people that depend on cassava for food and source of income. This project is part of the vision of the Next generation cassava project (NextGen Cassava) to deliver improved, higher yielding, broadly adapted, and stable genotypes at a much faster rate to farmers and cassava end-users.

Objectives

1. Calibration of the portable NIRS for rapid analyses of dry matter content (DMC) and carotenoids in cassava.
2. Genome-wide association and predictions of carotenoids.
3. Genome-enabled parentage assignment in a polycross nursery scheme.

REFERENCES

Anyanwu, C. N., Ibeto, C. N., Ezeoha, S. L., & Ogbuagu, N. J. (2015). Sustainability of cassava (Manihot esculenta Crantz) as industrial feedstock, energy and food crop in Nigeria. *Renewable Energy*, *81*, 745–752. http://doi.org/10.1016/j.renene.2015.03.075

Barratt, N., Chitundu, D., Dover, O., Elsinga, J., Eriksson, S., Guma, L., … Stevens, T. (2006). Cassava as drought insurance: Food security implications of cassava trials in Central Zambia. *Agrekon*, *45*(1), 106–123. http://doi.org/10.1080/03031853.2006.9523737

Belalcazar, J., Dufour, D., Andersson, M. S., Pizarro, M., Luna, J., Londoño, L., … Ceballos, H. (2016). High-throughput phenotyping and improvements in breeding cassava for increased carotenoids in the roots. *Crop Science*, *56*(6), 2916–2925. http://doi.org/10.2135/cropsci2015.11.0701

Boerner, V. (2017). On marker-based parentage verification via non-linear optimization. *Genetics Selection Evolution*, *49*(1), 50. http://doi.org/10.1186/s12711-017-0324-3

Büning-pfaue, H. (2001). Application of near infrared spectroscopy (NIRS) in the analysis of frying fats * Research Paper. *Symposium A Quarterly Journal In Modern Foreign Literatures*, *103*, 793–797.

Ceballos, H., Davrieux, F., Talsma, E. F., Belalcazar, J., Chavarriaga, P., & Andersson, M. S. (2017). Carotenoids in Cassava Roots. In *Carotenoids* (Vol. 3). http://doi.org/10.5772/intechopen.68279

Ceballos, H., Hershey, C., & Becerra-López-Lavalle, L. A. (2012). New Approaches to Cassava Breeding. *Plant Breeding Reviews*, *36*, 427–504. http://doi.org/10.1002/9781118358566.ch6

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

Chibueze Izah, S., Bassey, S. E., & Ohimain, E. I. (2018). Impacts of Cassava Mill Effluents in Nigeria. *Journal of Plant and Animal Ecology*. Retrieved from www.openaccesspub.org

Cock, J. H. (1987). Cassav: new potential for a neglected crop. *Field Crops Research*, *15*, 389–390. Retrieved from http://ciat-library.ciat.cgiar.org/Articulos_Ciat/Digital/SB211.C3_C5_C.2_Cassava_New_potential_for_a_neglected_crop.pdf

Davrieux, F., Dufour, D., Dardenne, P., Belalcazar, J., Pizarro, M., Luna, J., … Jaramillo, A. (2016). LOCAL regression algorithm improves near infrared spectroscopy predictions when the target constituent evolves in breeding populations. *Journal of Near Infrared Spectroscopy*, *24*(2), 109–117. http://doi.org/10.1255/jnirs.1213

De Alencar Figueiredo, L. F., Davrieux, F., Fliedel, G., Rami, J. F., Chantereau, J., Deu, M., … Mestres, C. (2006). Development of NIRS equations for food grain quality traits through exploitation of a core collection of cultivated sorghum. *Journal of Agricultural and Food Chemistry*, *54*(22), 8501–8509. http://doi.org/10.1021/jf061054g

Egesi, C. N., Ilona, P., Ogbe, F. O., Akoroda, M., & Dixon, A. (2007). Genetic variation and genotype X environment interaction for yield and other agronomic traits in cassava in Nigeria. In *Agronomy Journal* (Vol. 99, pp. 1137–1142). http://doi.org/10.2134/agronj2006.0291

El-Sharkawy, M. A. (2012). Stress-Tolerant Cassava: The Role of Integrative Ecophysiology-Breeding Research in Crop Improvement. *Open Journal of Soil Science*, *02*(02), 162–186. http://doi.org/10.4236/ojss.2012.22022

Esuma, W., Herselman, L., Labuschagne, M. T., Ramu, P., Lu, F., Baguma, Y., … Kawuki, R. S. (2016). Genome-wide association mapping of provitamin A carotenoid content in cassava. *Euphytica*, *212*(1), 97–110. http://doi.org/10.1007/s10681-016-1772-5

Fakir, M. S. A., Jannat, M., Mostafa, M. G., & Seal, H. (2012). Starch and flour extraction and nutrient composition of tuber in seven cassava accessions. *J. Bangladesh Agril. Univ*, *10*(2), 217–222. http://doi.org/http://dx.doi.org/10.3329/jbau.v10i2.14698

FAO. (2013). *Save and Grow: Cassava*. Retrieved from http://www.fao.org/3/a-i3278e.pdf

Goddard, M. E., & Hayes, B. J. (2007). Genomic selection. *Journal of Animal Breeding and Genetics*, *124*(6), 323–330. http://doi.org/10.1111/j.1439-0388.2007.00702.x

Halsey, M. E., Olsen, K. M., Taylor, N. J., & Chavarriaga-Aguirre, P. (2008). Reproductive biology of cassava (Manihot esculenta Crantz) and isolation of experimental field trials. *Crop Science*. http://doi.org/10.2135/cropsci2007.05.0279

Heaton, M. P., Leymaster, K. A., Kalbfleisch, T. S., Kijas, J. W., Clarke, S. M., McEwan, J., … Chitko-Mckown, C. G. (2014). SNPs for parentage testing and traceability in globally diverse breeds of sheep. *PLoS ONE*, *9*(4). http://doi.org/10.1371/journal.pone.0094851

Henry, G., Westby, A., & Collinson, C. (1998). European Group on RTB-Global Cassava End-uses & Markets, Phase 1-(FAO-ESCB) GLOBAL CASSAVA END-USES AND MARKETS: CURRENT SITUATION AND RECOMMENDATIONS FOR FURTHER STUDY. Retrieved from http://www.hubrural.org/IMG/pdf/global_cassava_end_use_study.pdf

Heslot, N., Yang, H.-P., Sorrells, M. E., & Jannink, J.-L. (2012). Genomic Selection in Plant Breeding: A Comparison of Models. *Crop Sci.*, *52*(1), 146–160. http://doi.org/10.2135/cropsci2011.06.0297

Iglesias, C., Mayer, J., Chavez, L., & Calle, F. (1997). Genetic potential and stability of carotene content in cassava roots. *Euphytica*, *94*(3), 367–373. http://doi.org/10.1023/A:1002962108315

Iwata, H., Hayashi, T., Terakami, S., Takada, N., Sawamura, Y., & Yamamoto, T. (2013). Potential assessment of genome-wide association study and genomic selection in Japanese pear Pyrus pyrifolia. *Breeding Science*, *63*(1), 125–140. http://doi.org/10.1270/jsbbs.63.125

Jansson C, Westerbergh A, Zhang J, Hu X, S. C. (2009). Cassava,a potential biofuel crop in the People's Republic of China. *Appl Energy*, *86*(595–599). Retrieved

from https://ac-els-cdn-com.proxy.library.cornell.edu/S0306261909002049/1-s2.0-S0306261909002049-main.pdf?_tid=40efbcbf-e6d5-477b-a26b-92d54e8ecc9d&acdnat=1530745177_3fdeb1e5fd9ca7e808aba0aed0a7165f

Jennings, D. L., & Iglesias, C. (2002). *Cassava: biology, production and utilization*. *Cassava: biology, production and utilization*. Wallingford: CABI. http://doi.org/10.1079/9780851995243.0000

Kawano, K. (2003). Thirty Years of Cassava Breeding for Productivity—Biological and Social Factors for Success. *Crop Science*, *43*(4), 1325. http://doi.org/10.2135/cropsci2003.1325

Kawano, K., Fukuda, W. M. G., & Cenpukdee, U. (1987). Genetic and Environmental Effects on Dry Matter Content of Cassava Root1. *Crop Science*, *27*(1), 69. http://doi.org/10.2135/cropsci1987.0011183X002700010018x

Lebot, V., Champagne, A., Malapa, R., & Shiley, D. (2009). NIR determination of major constituents in tropical root and tuber crop flours. *Journal of Agricultural and Food Chemistry*, *57*(22), 10539–10547. http://doi.org/10.1021/jf902675n

Lopez, A., Arazuri, S., Garcia, I., Mangado, J., Jaren, C., & Accepted, J. (2013). Review A REVIEW ON THE APPLICATION OF NEAR-INFRARED SPECTROSCOPY FOR THE ANALYSIS OF POTATOES FOR THE ANALYSIS OF POTATOES. http://doi.org/10.1021/jf401292j

Marten, G., Shenk, J., & Barton, F. (1989). Near infrared reflectance spectroscopy (NIRS): analysis of forage quality. *U.S. Department of Agriculture, Agriculture Handbook*, *643*, 1–110.

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829. http://doi.org/11290733

Nair, A. G. H., Sreekumar, J., Sheela, M. N., & Mohan, C. (2015). Influence of Different Pre-treatments on Enhancing Seed Germination in Cassava. *Journal of Root Crops*, *41*(2), 28–35. Retrieved from http://isrc.in/ojs/index.php/jrc/article/view/366/268

Ojulong, H., Labuschangne, M. T., Fregene, M., & Herselman, L. (2008). A cassava clonal evaluation trial based on a new cassava breeding scheme. *Euphytica*, *160*(1), 119–129. http://doi.org/10.1007/s10681-007-9590-4

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

Roggo, Y., Chalus, P., Maurer, L., Lema-Martinez, C., Edmond, A., & Jent, N. (2007). A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. *Journal of Pharmaceutical and Biomedical Analysis*, *44*(3 SPEC. ISS.), 683–700. http://doi.org/10.1016/j.jpba.2007.03.023

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

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redoña, E., … McCouch,

S. R. (2015). Genomic selection and association mapping in rice (Oryza sativa): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PLoS Genetics*, *11*(2), e1004982. http://doi.org/10.1371/journal.pgen.1004982

Teye, E., Huang, X., & Afoakwa, N. (2013). Review on the Potential Use of Near Infrared Spectroscopy (NIRS) for the Measurement of Chemical Residues in Food. *American Journal of Food Science and Technology*, *1*(1), 1–8. http://doi.org/10.12691/ajfst-1-1-1

United Nations, F. and A. O. (2015). FAOSTAT. Retrieved from http://faostat.fao.org/

Van Eenennaam, A. L., Weaber, R. L., Drake, D. J., Penedo, M. C. T., Quaas, R. L., Garrick, D. J., & Pollak, E. J. (2007). DNA-based paternity analysis and genetic evaluation in a large, commercial cattle ranch setting. *Journal of Animal Science*, *85*(12), 3159–3169. http://doi.org/10.2527/jas.2007-0284

Wolfe, M. D., Rabbi, I. Y., Egesi, C., Hamblin, M., Kawuki, R., Kulakow, P., … Jannink, J. (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), 0. http://doi.org/10.3835/plantgenome2015.11.0118

Zhang, X., Li, W., Yin, B., Chen, W., Kelly, D. P., Wang, X., … Du, Y. (2013). Improvement of near infrared spectroscopic (NIRS) analysis of caffeine in roasted arabica coffee by variable selection method of stability competitive adaptive reweighted sampling (SCARS). *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, *114*, 350–356. http://doi.org/10.1016/j.saa.2013.05.053

CHAPTER 2: RAPID ANALYSES OF DRY MATTER CONTENT AND CAROTENOIDS IN FRESH CASSAVA ROOTS USING A PORTABLE VISIBLE AND NEAR INFRARED SPECTROMETER (VIS/NIRS)

Abstract:

Portable Vis/NIRS are flexible tools for fast and unbiased analyses of constituents with minimal sample preparation. This study developed calibration models for dry matter content (DMC) and carotenoids in fresh cassava roots using a portable Vis/NIRS system. We examined the effects of eight data pre-treatment combinations on calibration models and assessed calibrations on processed and intact root samples. We compared Vis/NIRS derived-DMC to other phenotyping methods. The results of the study showed that the combination of standard normal variate and de-trend (SNVD) with first derivative calculated on two data points and no smoothing (SNVD+1111) was adequate for a robust model. Calibration performance was higher with processed than with intact root samples for all the traits although intact root models for some traits, especially total carotenoid content (TCC) (R^2_c = 96%, R^2_{cv} = 90%, RPD = 3.6 and SECV = 0.63), were sufficient for screening purposes. Using three key quality traits as templates, we developed models with processed fresh root samples. Robust calibrations were established for DMC (R^2_c = 99%, R^2_{cv} = 95%, RPD = 4.5 and SECV = 0.9), TCC (R^2_c = 99%, R^2_{cv} = 91%, RPD = 3.5 and SECV = 2.1) and all Trans β-carotene (ATBC) (R^2_c = 98%, R^2_{cv} = 91%, RPD = 3.5 and SECV = 1.6). Coefficient of determination on independent validation set (R^2_p) for these traits were also satisfactory for ATBC (91%), TCC (88%) and DMC (80%). Compared to other methods, Vis/NIRS-derived DMC from both intact and processed roots had higher correlation (>0.95) with the ideal oven-drying than from specific gravity method (0.49). There was equally a high correlation (0.94) between the intact and processed Vis/NIRS DMC. Therefore, the portable Vis/NIRS could be employed for the rapid analyses of DMC and quantification of carotenoids in cassava for nutritional and breeding purposes.

Keywords: Cassava, near infrared spectroscopy, Calibration, Partial Least Square Regression, Dry matter content, Carotenoids.

Introduction:

Near infra-red spectroscopy (NIRS) is one of the most important analytical techniques based on the vibrational properties of atoms in molecules (Lopez et al. 2013; Stuart 2004). NIRS has gained wide application over years in the analyses of many materials including agricultural and food products (Manley 2014; dos Santos et al. 2013). When compared to other analytical and chemical methods, NIRS offers a fast, non-destructive

alternative for the simultaneous analyses of many constituents (Büning-Pfaue 2003). It requires minimal to no sample preparation, and it is economically efficient and non-hazardous to the environment (Guoquan Lu, Huang, and Zhang 2006).

Near infra-red spectroscopy is an ideal phenotyping tool in plant breeding, particularly in this era when new breeding techniques are being adopted (Cabrera-Bosquet et al. 2012; Jannink, Lorenz, and Iwata 2010), requiring the phenotyping of thousands of individuals at low cost and with high precision and speed. NIRS permits the timely screening of many samples and variables that would have been too expensive to assay by other analytical methods (Jannink, Lorenz, and Iwata 2010; Guo-quan Lu, Huang, and Zhang 2006). One of its notable advantages, is its ability to measure samples in different states – in solid and liquid forms (Blanco and Villarroya 2002).

Breakthroughs in technology have led to the increasing availability of spectrophotometers of various ranges in a portable format, and this provides greater flexibility for field-based analyses of constituents. The portable NIRS, in some cases covering both the visible and near infrared regions (Vis/NIRS), has the advantage of further reducing the need for sample transportation to a laboratory and processing. It provides a quality phenotyping method for breeding programs, especially where standard laboratories are not available or their operation is hampered by factors such as poor infrastructure and lack of highly skilled experts. It is believed (Guoquan Lu, Huang, and Zhang 2006) that over the long-term developing NIRS is cheaper than the establishment of many protocols for laboratory analyses of different traits, which in

most cases are slow, costly and impractical for large-scale screening in plant breeding and nutritional quality analyses (Cabrera-Bosquet et al. 2012; Sánchez et al. 2014).

In cassava breeding, the adoption of new methods has necessitated standardized and accurate phenotyping tools for efficient improvement, especially for complex traits (Hernán Ceballos et al. 2015). Availability of phenotyping tools for accurate and large scale screening of materials, particularly at early stages of cassava breeding, will reduce the loss of important genetic information and facilitate the breeding of end-user and farmer-preferred cultivars (Asrat et al. 2010). The current phenotyping techniques for some key traits are laborious and time-consuming for large-scale screenings. Estimates could be influenced by sampling and sample preparation including weight and number of roots used in the prevalent specific gravity method (Fukuda et al. 2010; Kawano, Fukuda, and Cenpukdee 1987; Pérez et al. 2011) and inconsistency of power supply in the oven-drying method. Similarly, carotenoid quantification using color intensity (Sánchez et al. 2006) could be subjective and inefficient in advanced populations of yellow genetic materials. Conversely, laboratory processes using high-performance liquid chromatography (HPLC) or UV-Visible spectrophotometer are low-throughput (less than 10 or 40 samples per day, respectively) (Sánchez et al. 2014).

The use of NIRS for the analyses of traits on fresh cassava roots have been previously reported (Sánchez et al. 2014) and has led to significant changes in a breeding system (Belalcazar, Dufour, Andersson, Pizarro, Luna, Londoño, Morante, Jaramillo, Pino, López-Lavalle, Davrieux, Talsma, and Ceballos 2016). However, these studies used a stationary tabletop NIRS device with processed root samples – peeled and mashed,

aimed at overcoming the reported uneven concentration of traits in cassava roots (Ortiz, Sánchez, and Morante 2011). Nevertheless, the possibility of reduced sample preparations using intact samples have been reported in other scenarios (Campbell et al. 1999; De Alencar Figueiredo et al. 2006; Arganosa et al. 2006). Preparation of cassava root samples before NIRS analysis adds to the harvesting time and the overall cost of phenotyping. The use of a full-range portable Vis/NIRS device has not been reported in cassava breeding and the possibility of reduced root processing has not equally been explored. Obtaining a good relationship between calibrations from processed and intact samples could enable simultaneous field-based screening of materials on various important traits and the overall reduction of phenotyping cost.

Generally, when working with NIRS, the spectral variation of interest can be masked by additive and/or multiplicative light scattering, background noise and baseline drifts arising from differences in particle sizes and effective path-length (Pizarro et al. 2004; Rinnan, Berg, and Engelsen 2009). It is therefore important to adopt suitable data pre-processing methods to minimize the influence of these physical effects on the NIRS calibration (Rinnan, Berg, and Engelsen 2009; Blanco et al. 1997).

In this study, we assess the use of a portable Vis/NIRS device for the analysis of important fresh cassava quality traits on both processed (mashed) and non-mashed (intact) root samples. We assess the impact of data pre-processing for possible increase in the predictive ability of the calibration models. The ultimate goal of this study was to develop calibration models using the portable Vis/NIRS for the analyses of DMC and carotenoids in fresh cassava roots, which could accelerate accurate phenotyping and

general improvement of cassava. To examine the usefulness of this tool on dry matter quantification, we compared dry matter values derived from the conventional specific gravity method and predicted values from the portable Vis/NIRS (intact and mashed) to the ideal oven-drying method.

Materials and methods

***Calibration samples*:** In 2015, first calibration set (Table 2.1) was developed using clones (U15I, N = 113) from the germplasm collection of the National Root Crops Research Institute (NRCRI), Umudike, Nigeria. Single root samples were randomly selected from harvested clones of a training population (TP) established for the implementation of genomic selection. The selected roots were peeled and chopped into pieces (about 3x10 mm) using kitchen knives.

A second calibration set (Table 2.1) was developed in 2016 at the International Center for Tropical Agriculture (CIAT), Cali-Palmira, Colombia. Between two to three root samples were collected from F1 seedling plants of different half- and full-sib families of varying sizes (Belalcazar, Dufour, Andersson, Pizarro, Luna, Londoño, Morante, Jaramillo, Pino, López-Lavalle, Davrieux, Talsma, and Ceballos 2016; H. Ceballos et al. 2013). Additional clones with white parenchyma from the germplasm collection at CIAT were added in order to balance the calibration set. All the field sampling and selections were carried out early in the morning and the selection of individuals for carotenoid was based on yellow/orange color intensity of roots, which is closely associated with high carotenoids, especially TCC and total beta carotene (TBC) in cassava (Sánchez et al. 2006; Sánchez et al. 2014). The selected roots were peeled and

mashed into a homogenous sample in the laboratory using an Essen Skymsen food processor (Model: PA-7SE, Brusque, Brazil).

Third calibration set (Table 2.1) was developed in 2016 at NRCRI for DMC using intact and mashed root samples. Two or three roots were randomly selected from one or two plants in a plot of five plants per clone from the NRCRI TP. The selected roots were evaluated for DMC by specific gravity before peeling and mashing using a portable power-operated grater.

The 2016 set from NRCRI and a subset of the calibration set from CIAT were used for the comparison of calibrations from intact and mashed root samples.

Spectral data collection: A portable Vis/NIRS device (QualitySpec Trek: S-10016) was used to collect spectral data on both intact and mashed root samples. Spectral data on intact roots were obtained by placing roots in contact with the window of the portable Vis/NIRS device. Each spectrum collected is in fact the average of 50 spectra collected over a period of five seconds. Three spectra per root were taken respectively on the proximal, middle and distal regions of roots at NRCRI and CIAT in 2016. The selected root samples were first peeled, rinsed with water and dried with a paper towel before spectra collection. However, depending on the size of the roots, spectral data were only collected from the transverse section of the proximal end of the root and few samples from proximal and distal ends in 2015 at NRCRI. The mean spectrum for each sample was used for calibration.

For mashed samples, spectral data were collected from about 8g of homogenized mashed roots in quartz sampling cups placed against the window of the portable

Vis/NIRS device. Two replications were done per sample, and spectrum averages were used for analyses.

Reference/Wet chemistry:

Dry matter content (DMC): At both locations (CIAT and NRCRI), dry matter was measured as the percentage of dry weight relative to a given fresh weight of samples after oven-drying. Between 80 and 110 g (measured to 0.1 mg precision) of the mashed and homogenized roots were oven-dried at a constant temperature of 105^0C for 24 hours at CIAT. At NRCRI in 2015, 10 g of the chopped samples were weighed before and after oven-drying while in 2016, 20 g of the mashed samples were dried in two replications. The oven temperature at NRCRI was targeted for TTT^{0}C. Depending on the duration and source of power, samples were weighed after drying. The average DMC of the two replications was used for analyses. Specific gravity method (Fukuda et al. 2010) was carried out before the selected two or three roots were processed – peeled, washed and dried with a paper towel in 2016 at NRCRI.

Carotenoids: The reference samples at CIAT were measured for carotenoids using a HPLC system (Agilent Technologies 1200 series, Waldbronn, Germany). To avoid quality degradation of samples, an average of six (6) samples per day were analyzed with the HPLC. As previously described (Sánchez et al. 2014) and complying with the HarvestPlus standards for optimum carotenoids retention (D. . Rodriguez-Amaya and Kimura 2004), all the extractions were performed on fresh roots with minimal exposure to light, high temperatures and reduction of time between mashing and extraction. The HPLC reference traits included – TCC, all-trans β-carotene (ATBC), violaxanthin

(VIO), Lutein (LUT), 15-Cis beta-carotene (15CBC), 13-Cis beta-carotene (13CBC), Alpha carotene (AC), 9-Cis beta-carotene (9CBC) and phytoene (PHY).

Measurement of TCC at NRCRI in 2015 was carried out at the NRCRI Carotene laboratory in Umudike following the standard laboratory extraction method using acetone with mortar and pestle and spectrophotometric quantification as described in the Harvest-Plus handbook (D. . Rodriguez-Amaya and Kimura 2004). Homogenized samples of 10g were ground in a mortar with 3g of Hyflosuperce (Celite) and 50mL of cold acetone. The mixture was filtered with a Buchner funnel with filter paper while the mortar, pestle, funnel and residue were washed into a suction flask and observed to be sure that the washings or residue were devoid of color. Otherwise, the residue was returned to the mortar for further maceration, filtering and washing. The next step involved the petroleum ether partitioning where about 20mL of petroleum ether and acetone were added into a 500mL separator funnel with Teflon stop-cock. Distilled water (~300mL) was slowly added into the mixture. The two phases were allowed to separate and the lower, aqueous phase was discarded while the remaining phase was washed 3 to 4 times with distilled water (~200mL) to remove residual acetone. The petroleum ether phase was transferred into a 25mL volumetric flask through a funnel containing glass wool and anhydrous sodium sulphate (about 15 g) to remove the residual water. The absorbance of the extract was measured at 450 nm using a spectrophotometer (Electron Corporation Ltd – GENESYS 10 Series) and TCC was derived using:

$$TCC\ (\mu/g)\ = \frac{A\ \times\ Volume\ (mL)\ \times 10^4}{A_{1cm}^{1\%}\ \times\ sample\ weight\ (g)},$$

$where, A =$ absorbance; $Volume =$ total volume of extract; $A_{1cm}^{1\%}$ $=$ absorption coefficient of β − carotene in Petroleum ether (equals 2592).

Data pre-processing and model development:

Prior to model development, spectral data were first transformed to log (1/R) using ViewSpec Pro software (ASD 2008), and the full Vis/NIRS wavelength range (350 – 2500nm) was subjected to pre-treatments for the correction of interferences on three segments of the wavelengths (350nm -1000nm, 1001nm – 1800nm and 1801nm – 2500nm). The effect of two light-scatter correction methods - Standard Normal Variate and De-trending (SNVD) (Barnes, Dhanoa, and Lister 1989) and Multiplicative Scatter Correction (MSC) (Geladi, MacDougall, and Martens 1985) were tested on four derivative and smoothing options. The options are given by four digits (D, G, S1, S2): where D indicates the derivative order number (0 indicates no derivation, 1 means the first derivative, and so on), G indicates the gap (the number of data points over which derivation is computed), S1 indicates the number of data points in the first smoothing (1 means no smoothing) and S2 indicates the number of data points in the second smoothing, where 1 means no smoothing. The eight pre-treatment methods (SNVD+1111, SNVD+2111, SNVD+1551, SNVD+2551, MSC+1111, MSC+2111, MSC+1551 and MSC+2551) were compared to no treatment in each calibration set for DMC, TCC and ATBC.

SNVD: The SNVD correction requires two algorithms that are usually applied together. The first algorithm is the Standard Normal Variate (SNV) and is used for correcting scattering when the effective path length and baseline varies among samples of a data set (Pizarro et al. 2004) and for granular or powdery samples or when the particle sizes vary among samples (Barnes, Dhanoa, and Lister 1989). SNV is usually applied first to correct the effects of the multiplicative interferences of scatter and particle size differences by removing the mean and scaling to unit variance. SNV correction is given by:

$S_i = {(S_0 - S_v)}/{S_d}$, where S_i = corrected spectrum,

S_o = original individual spectrum measured by the NIR device,

S_v = average value of the sample spectrum to be corrected, S_d = standard deviation of the sample spectrum.

De-trending attempts to remove the additional variation in baseline shift and curvelinearity by fitting the spectral values of a given *i* spectrum at *k* wavelength ($S_{i,k}$) to a polynomial function – for example, a quadratic function $(`S_{i,k})$ (Di) and subtracts the function (quadratic baseline) from the spectral values (Dii) (Blanco et al. 1997):

$$`S_{i,}k = a + b.k + c.k^2 \quad \ldots\ldots \quad Di$$

$$S_{i,\,k(De-trend)} = Si,\ k - `Si,\ k \quad \ldots\ldots\ldots \quad Dii$$

SNVD does not require external references and each spectrum is treated independently of others in the training set (Rinnan, Berg, and Engelsen 2009).

MSC: This method attempts to correct for particle size dependence by linearizing each spectrum to an ideal or reference sample spectrum which in most cases is the average spectrum obtained from all the data in the training set. The slope and offset of the sample spectra are adjusted to the ideal average spectra to give the MSC corrected spectrum (Geladi, MacDougall, and Martens 1985; Rinnan, Berg, and Engelsen 2009). The process of MSC correction, assuming the reference is the mean, includes:

a. Reference spectrum calculation: $\bar{S}_j = \sum_{i=1}^{n}(S_{i,j})/n$
b. Using spectral responses in each spectrum to calculate a linear regression against the corresponding points in the reference spectrum: $S_i = a_i\bar{S} + b_i$
c. Subtracting the slope from the regression on the original spectrum and dividing with the offset values to obtain MSC corrected spectrum:

$$S_{i(MSC)} = (S_j - b_i)/\, a_i,$$

where, S = spectral responses for all the wavelengths; $\bar{S}$ = average responses of all the training set spectra at each wavelength; S_i = responses for a single spectrum in the training set; n = number of training spectra; a_i and b_i = slope and offset coefficients of the linear regression of the mean spectrum vector $\bar{S}$ versus S_j spectrum.

Derivatives and Smoothing: The basic method of derivation is finite difference where: the first-order derivation takes the difference between two values with a given gap size while second order derivative is then estimated by calculating the difference between two successive points of the first-order derivative spectra (Rinnan, Berg, and Engelsen 2009; D. Li et al. 2011). In place of the basic derivative which is usually not feasible for most real measurements due to noise inflation, the modified smoothing and derivative of the Norris-Williams approach (Rinnan, Berg, and Engelsen 2009) is usually the preferred option:

a. Smooth the spectra. Average over a given number of points.

$x_{smooth,i} = \frac{\sum_{j=-m}^{m} x_{org,i+j}}{2m+1}$, where m is the radius of the smoothing window centered on the current measurement point i.

b. Derive at each wavelength. For the first derivative take the difference between two smoothed values at a given gap distance and for the second-order derivative, take twice the smoothed value at point i and the smoothed value at a gap distance on either side:

$x_i' = x_{smooth,i+gap} - x_{smooth,i-gap}$

$x_i'' = x_{smooth,i-gap} - 2.x_{smooth,i} + x_{smooth,i+gap}$.

Spectra pre-treatments as well as model development were implemented in Win-ISI 4.0 software (Infrasoft International and FOSS, Hillerod, Denmark). The modified Partial Least Squares (MPLS) algorithm was used to set up a multivariate model based on the reference chemical values and the pre-treated spectra. The MPLS is a partial least square regression (PLSR), modified to scale the reference data and reflectance data at each wavelength to have a standard deviation of 1.0 (Marten, Shenk, and Barton 1989; Shenk and Westerhaus 1991). It reduces the spectral data to a few orthogonal combinations (or factors) of absorbance that account jointly for the most spectral and reference value information (Freschet et al. 2011).

Validation of models:

Models were developed using individual calibration sets across locations and years and each model was used to predict the values of other sets on either the mashed or intact root sample categories. However, because of the differences in references value standards, the major calibration set from mashed samples developed at CIAT was

divided into two - calibration and validation sets (Table 2.2) using the *naes* calibration sampling algorithm (Naes et al. 2002) - *prospectr* package (Stevens and Ramirez Lopez 2013) in R for model development and validation. The *naes* sampling procedure usually uses cluster analysis to select calibration samples from large multivariate datasets. By retaining principal components explaining at least 99 percent of the total variance following a PCA on the spectral variables, k-means clustering (1000 iterations) was carried out on the principal component scores, with a number of clusters equal to the number of desired calibration samples (Table 2.2). The calibration set was constituted by drawing samples from the center of each cluster, leaving the remaining samples as validation set. This systematic sampling approach was used to ensure that the calibration set was representative of the dataset than a random sampling. The calibration set from intact roots in CIAT had small sample size and was only used to evaluate the possibility of direct unprocessed root assay. In order to perform cross-predictions in the WinISI software, the ASD spectra (350nm – 2500nm in 1nm gap) were trimmed to a range (400nm – 2500nm in 2nm gaps) compatible with the Win-ISI software.

Reported calibration statistics included the standard deviation (SD), coefficient of determination (R^2), standard error of calibration (SEC) and standard error of cross-validation (SECV). In each model, leave-one-out cross-validation (iteratively removing one sample and predicting it using the remaining samples) was used for internal model assessment. The optimum number of PLS latent variables, which maximizes the covariance between the response and predictor variables was selected based on the minimum value of SECV. In addition, the ratio of performance to deviation (RPD = SD/SECV) as well as standard error of prediction (SEP) and standard error of prediction

corrected for bias [SEP (C)] were used to evaluate the quality of the prediction models (Williams and Sobering 1993; Sánchez et al. 2014). Unlike SEC and SECV, RPD is independent of parameter units and can therefore be compared between parameters (Davrieux et al. 2016).

Samples whose spectra had high Mahalanobis distance (H-outliers) with reference to the average spectrum or for which the difference between the reference and the predicted value was much higher than the standard error of cross-validation (SECV) (t-Outliers) were defined as outliers and removed in the calibration model. As suggested by (Tillmann, Reinhardt, and Paul 2000; Terhoeven-Urselmans 2007), the outlier limits were set to 10 (H-outliers) and 2.5 (t-outliers). Up to three iterations of outlier identification and re-calibration (Wang et al. 2017) were allowed (Sánchez et al. 2014; Davrieux et al. 2016; Shenk and Westerhaus 1991). Some of the models were stable (no outliers detected) after one or two iterations.

Correlation of DMC from different methods: To assess the relevance of the Vis/NIRS-derived DMC relative to the standard oven-drying and the conventional gravitational methods, we compared the Vis/NIRS-derived values from mashed and intact sampling with DMC from oven drying and specific gravity methods from 173 samples at NRCRI in 2016. The oven drying DMC has been described above. Specific gravity DMC is derived from the linear relationship between DMC and specific gravity (SG):

DMC = 158.3SG – 142, where SG is the ratio of weight of the sample in air to the difference between weights of the sample in air versus water.

The Pearson correlation was used to assess the relationships among the four various DMC sets – oven drying, SG-derived, mashed NIRS-derived and intact root NIRS-derived DMC. The regression between specific gravity and DMC for the selected samples was also estimated (Kawano, Fukuda, and Cenpukdee 1987; Fukuda et al. 2010).

Results and Discussion

Statistics of reference data:

It is important to ensure adequate range and precision of traits in developing NIRS calibrations (Fox et al. 2012). The range of the reference values for DMC on both sampling methods - intact and mashed roots was between 16% and 51% which seems applicable to many breeding programs for immediate evaluations and feasible DMC improvement (Table 2.1). The mean DMC at Umudike in 2015 on intact root samples (U15I) was higher than the mean of the reference data for the same trait generated at CIAT in 2016 on intact root samples (C16I) but lower than what was obtained at Umudike in 2016 on both intact and mashed (U16I/M) root samples (Table 2.1). The DMC of the intact/mashed (U16I/M) set from NRCRI in 2016 however, was higher than mashed samples from CIAT (C16M). The quantification approaches for TCC were different at NRCRI and CIAT but the mean TCC at CIAT was higher (17.95µg $g^{-1}$and 14.91µg g^{-1}on intact and mashed root samples, respectively) than NRCRI (2.14µg g^{-1}) from only intact root samples. Varying ranges of carotenoids were obtained from the HPLC analyses for the carotenoids, although TCC and ATBC were used for most of the carotenoid analyses.

Table 2.1: Description of calibration sets developed at NRCRI Umudike, Nigeria and CIAT, Cali Colombia in 2015 and 2016 on intact and mashed root samples.

Statistics	U15I	C16I	U16I/M	C16M	U15I	C16I	C16M	C16M
	DMC (%)				TCC (µg g^{-1})			ATBC (µg g^{-1})
No.	113	66	194	173	113	65	173	173
Mean	35.75	20.14	38.52	36.16	2.61	17.95	14.91	10.07
SD	7.95	4.27	5.76	4.16	2.14	3.84	7.73	5.86
Min.	16.34	16.54	16.47	20.14	0.10	10.09	0.70	0.03
Max.	50.98	41.98	50.00	44.13	8.82	26.15	30.84	21.02

U15I = Calibration set on intact root samples at Umudike in 2015; U16I/M = Calibration set on intact and mashed roots at Umudike in 2016; C16I = Calibration set on intact roots from CIAT in 2016; C16M = Calibration set on mashed roots from CIAT in 2016. Carotenoids (TCC and ATBC) data are on a fresh weight basis.

The use of the *naes* sampling algorithm enabled an even distribution of the calibration and validation sets of the mashed samples developed at CIAT in 2016 as seen in their descriptive statistics – mean, standard deviation and range (Table 2.2).

Table 2.2: Descriptive statistics for model calibrations and independent set validations for DMC, TCC and ATBC using mashed root samples from CIAT, 2016.

Traits	Calibration set					Validation set				
	No.	Mean	SD	Min.	Max.	No.	Mean	SD	Min.	Max.
DMC	120	36.06	4.31	20.14	43.30	53	36.40	3.84	27.35	44.13
TCC	119	14.94	7.87	1.00	30.84	54	14.85	7.49	0.70	26.15
ATBC	119	9.97	5.89	0.029	21.02	54	10.29	5.85	0.31	20.33

Effect of pre-processing methods on calibration statistics for different calibration sets on intact and mashed root samples:

Much emphasis has been laid on the need for optimum mathematical pre-treatment of spectra prior to model generation in order to minimize the impact of interferences arising from variation in particle sizes, optical path-length and crystalline forms on spectra (Roggo et al. 2007). Given that the portable Vis/NIRS has not been used in trait analyses in cassava, several pre-treatment combinations were tested in order to identify the best combination that would minimize the effect of interferences on prediction. A total of eight pre-processing combinations were assessed on the different calibration sets for different traits and from the two sampling methods – intact and mashed samples. The reported performances of the eight pre-treatment methods are based on R^2 values for calibration (R^2_c) and cross-validation (R^2_{cv}) (Table 2.3). Usually, R^2 of 0.50 has been classified as useful in the discrimination of concentrations, between 0.60-0.82 for

screening and quantification, 0.83-0.90 is important in most applications, 0.92-0.96 is useful in most applications especially in quality assurance and above 0.98 is important for all applications (Fox et al. 2012). Also, RPD has been used in evaluating the robustness of a model. RPD values greater than three (>3.0) has been considered sufficient; 2.0-3.0 (good); 1.5-2.0 (medium) and less than 1.5 (poor) for analytical quality in various applications (Wang et al. 2017; D'Acqui, Pucci, and Janik 2010; Williams and Sobering 1993).

The average R^2_c and R^2_{cv} for DMC across the different calibration sets showed that SNVD+1111 had the highest average R^2_c (94%) and R^2_{cv} (73%), slightly higher than MSC+1111 with average R^2_c of 92% and R^2_{cv} of 72% (Table 2.3). The average R^2_c from SNVD+1111 was also higher (95%) than MSC+1111 (94%) although the R^2_{cv} using MSC+1111 (86%) was higher than that of SNVD+1111 (83%) for TCC calibrations. The highest average R^2_c (~100%) for ATBC was obtained from MSC+1111 and MSC+2551 whereas the highest R^2_{cv} (~95%) was obtained from SNVD (1111 and 2551) and MSC+2551. Across the three traits, overall average performance from SNVD+1111 (R^2_c = 95% and R^2_{cv} = 79%) and MSC+1111 (R^2_c = 94% and R^2_{cv} = 78%) were higher than other pre-treatments. It was observed that R^2_c and R^2_{cv} from other pre-treatment methods on individual sets were in some cases similar or even greater than values from SNVD+1111 or MSC+1111 but in all cases, performance from SNVD+1111 was still relatively high.

Compared to the no pre-treatment, the number of independent variables (spectra) used in pre-treatment evaluations often varied with the treatment methods. The average R^2_c

and R^2_{cv} values from no pre-treatment for DMC and TCC calibrations were lower than the best pre-treatments from SNVD+1111 and MSC+1111. However, the R^2_{cv} on individual calibration sets from no pre-treatment especially with the calibration set from CIAT in 2016 (C16M) was in some cases, higher than the R^2_{cv} from any of the pre-treatment methods. For example, the highest average R^2_{cv} (97%) for ATBC was obtained from no pre-treatment.

Percentage improvement of models arising from pre-treatments was higher using intact than mashed root samples. This could be attributed to higher levels of interference when using intact root than mashed samples.

Therefore, based on the R^2_c and R^2_{cv} performances, it seemed that the most promising pre-treatment using the Vis/NIRS device was SNVD+1111. The high performance of SNVD has been previously reported (Barnes, Dhanoa, and Lister 1989) for the same traits in cassava although using a different instrument and on different derivative and smoothing gaps (2,5,5,1) (Davrieux et al. 2016; Sánchez et al. 2014). It is therefore necessary to adopt the most promising pre-treatment when working with NIRS devices.

Table 2.3: The effect of mathematical pre-treatments on models from different calibration sets.

Pre-trmt.	Der.& Sm.	R^2	U15I	U16I	C16I	U16M	C16M		U15I	C16I	C16M		C16M
			DMC (%)					AV. DMC (%)	TCC (µg)			Av. TCC (µg)	ATBC (µg)
NONE	0,0,1,1	R^2_c	0.66	0.70	0.54	0.83	0.96	0.74	0.94	0.52	0.97	0.81	0.970
		R^2_{cv}	0.55	0.64	0.44	0.79	0.96	0.68	0.91	0.40	0.96	0.76	**0.970**
SNVD	1,1,1,1	R^2_c	0.91	0.90	0.96	0.96	0.99	**0.94**	0.96	0.90	0.99	**0.95**	0.987
		R^2_{cv}	0.64	0.65	0.55	0.84	0.95	**0.73**	0.90	0.67	0.93	0.83	0.945
	1,5,5,1	R^2_c	0.80	0.92	0.60	0.84	0.97	0.83	0.94	0.95	0.96	**0.95**	0.952
		R^2_{cv}	0.64	0.73	0.50	0.80	0.95	0.72	0.90	0.76	0.93	**0.86**	0.928
	2,1,1,1	R^2_c	0.81	0.85	0.60	0.93	0.97	0.83	0.91	0.89	0.98	0.93	0.982
		R^2_{cv}	0.41	0.37	0.22	0.46	0.55	0.40	0.57	0.62	0.86	0.68	0.847
	2,5,5,1	R^2_c	0.79	0.86	0.64	0.87	0.97	0.83	0.95	0.84	0.96	0.92	0.994
		R^2_{cv}	0.55	0.64	0.48	0.80	0.95	0.68	0.84	0.61	0.92	0.79	0.947
MSC`	1,1,1,1	R^2_c	0.77	0.91	0.97	0.96	0.99	0.92	0.95	0.89	0.99	0.94	**0.995**
		R^2_{cv}	0.59	0.68	0.57	0.83	0.95	0.72	0.89	0.64	0.94	0.82	0.944
	1,5,5,1	R^2_c	0.78	0.91	0.75	0.87	0.97	0.86	0.94	0.93	0.95	0.94	0.947
		R^2_{cv}	0.60	0.74	0.53	0.80	0.95	0.72	0.89	0.68	0.92	0.83	0.924
	2,1,1,1	R^2_c	0.79	0.85	0.60	0.93	0.97	0.83	0.90	0.89	0.99	0.93	0.984
		R^2_{cv}	0.41	0.41	0.22	0.46	0.56	0.41	0.57	0.62	0.86	0.68	0.852
	2,5,5,1	R^2_c	0.79	0.86	0.64	0.87	0.97	0.83	0.95	0.84	0.96	0.92	**0.995**
		R^2_{cv}	0.58	0.64	0.48	0.80	0.95	0.69	0.85	0.61	0.92	0.79	0.951

U15I = Calibration set on intact root samples at Umudike in 2015; U16I = Calibration set on intact root samples at Umudike in 2016; U16M = Calibration set on mashed root at Umudike in 2016; C16I = Calibration set on intact roots from CIAT in 2016; C16M = Calibration set on mashed roots from CIAT in 2016.

Calibration models on intact and mashed root samples:

Given the higher measurement speed and minimum processing of root samples using intact roots, this method would be highly desirable with acceptable model performance. Higher accuracies with ground/processed samples have been obtained in similar settings in peas (Arganosa et al. 2006), grains and seeds (Williams and Sobering 1993; Campbell et al. 1999) and the correlation between predictions from intact and ground samples could be high enough for routine screening purposes (Campbell et al. 1999; Arganosa et al. 2006).

Using RPD as a calibration statistic to assess models developed from mashed and intact roots, the result showed that the RPD values for DMC from mashed samples were 2.50 and 4.32 from U16M and C16M calibrations, respectively (Table 2.4a). The RPD from intact root samples on both years – 2015 and 2016 at Umudike was 1.68 (Table 2.4b). For better comparison using the same number of clones from CIAT in 2016 from the mashed samples (C16M66) and intact samples (C16I66), the calibration from mashed samples was evidently higher than that of intact root samples (Table 2.4a and 2.4b). Similar results were obtained when using the same number of samples from NRCRI in 2016 (Result not presented). However, the R^2_c of models from intact roots were still high (>86%) with R^2_{cv} ranging from 55% to 65% (Table 2.4b).

Table 2.4: Calibration assessments of DMC from different calibration sets on mashed (a.) and intact (b.) root samples for DMC.

Calibration set	SEC	R^2_c	SECV	R^2_{cv}	SD	RPD
a. Calibrations of DMC on mashed root samples						
U16M	0.91	0.96	1.87	0.84	4.67	2.50
C16M	0.41	0.99	0.95	0.95	4.10	4.32
C16M66	0.52	0.99	1.04	0.94	4.24	4.08
b. Calibrations of DMC on intact root samples						
U15I	2.16	0.91	4.37	0.64	7.34	1.68
U16I	1.78	0.86	2.80	0.64	4.71	1.68
C16I66	0.77	0.96	2.59	0.55	3.86	1.49

The calibration performance for carotenoids showed that the R^2_c for most of the carotenoids was 99% except in alpha-carotene (80%), lutein (88%), phytoene (91%) and violin (94%), which are found at low concentrations in cassava roots (Appendix 2.1). However, the R^2_{cv} for these traits varied from 41% in phytoene to 95% in ATBC (Appendix 2.1 and Table 2.5 respectively). Similar to the R^2_{cv}, the RPD was lowest in phytoene (1.31) and highest in ATBC (4.29). Comparing TCC calibration from mashed root at CIAT to TCC from intact root at NRCRI in 2015, both calibrations had very

good calibration performances (Table 2.5a and 2.5b) (Figures 2.1 and 2.2). However, the calibration performance from C16M (R^2_c = 99% and R^2_{cv} = 93%; RPD = 3.79) was higher than U15I (R^2_c = 96% and R^2_{cv} = 90%; RPD = 3.16).

Table 2.5: Calibration assessments of Carotenoids from mashed (a.) and intact (b.) root samples.

Cal. set	Traits (µg)	No.	Range	Mean	SD	SEC	R^2_c	SECV	R^2_{cv}	RPD
a. Calibration for carotenoids from mashed samples using the entire calibration set from CIAT.										
C16M	TCC	164	0.70-28.87	14.84	7.32	0.64	0.99	1.93	0.93	3.79
	ATBC	161	0.03-20.33	10.05	5.53	0.64	0.99	1.29	0.95	4.29
b. Calibration for TCC using intact root samples from Umudike in 2015.										
U15I	TCC	102	0.10-8.82	2.45	1.99	0.38	0.96	0.63	0.9	3.16

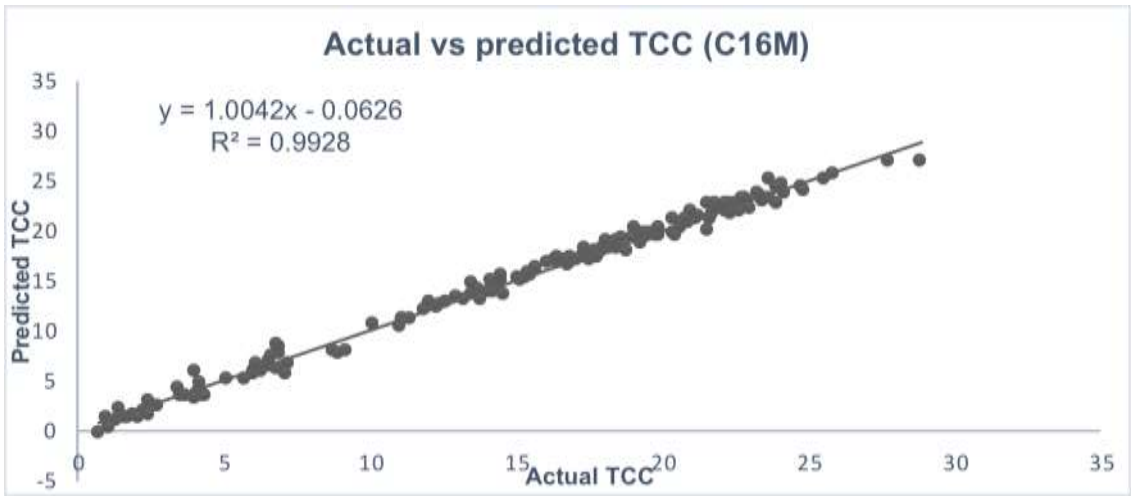

***Figure 2.1. Calibration for TCC on mashed samples using data from CIAT in* 2016.**

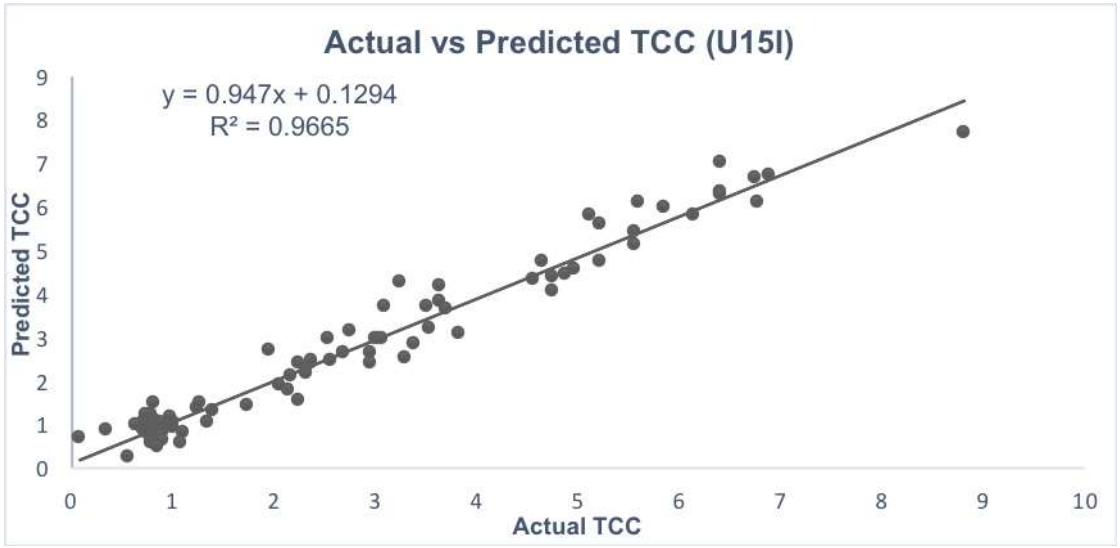

Figure 2.2. Calibration for TCC on intact samples using data from NRCRI in 2015.

Similar to results obtained for DMC calibration using the same number of individuals for comparison between calibrations from mashed and intact root samples, the calibration statistics for carotenoids from mashed calibrations were still better than the calibrations from intact root (Table 2.6a, 2.6b and Appendix 2.2). The R^2_c from mashed

samples varied between 73% and 99% while intact root calibrations were greater than 67% except in an extreme case where lutein was less than 50%. The R^2_{cv} varied from 33% to 93% in mashed calibrations and 10% to 81% in intact root calibrations. Various RPD values were obtained from the two sampling methods with values from mashed roots still higher than that from intact root calibrations.

Higher prediction models from ground against whole or intact samples have been reported (Arganosa et al. 2006; De Alencar Figueiredo et al. 2006; Williams and Sobering 1993) and could be attributed to higher scattering noise for spectra obtained from intact samples (De Alencar Figueiredo et al. 2006) but the correlations between derived values from ground and intact samples are usually high (De Alencar Figueiredo et al. 2006; Arganosa et al. 2006). The difference between calibrations from the two sampling methods are minimal with small and less heterogeneous grains (De Alencar Figueiredo et al. 2006). This means that reducing interferences and heterogeneity in the case of cassava (Ortiz, Sánchez, and Morante 2011), could reasonably improve accuracy from intact samples.

Table 2.6: Carotenoids calibrations from mashed (a) and intact (b) root samples from CIAT using the same sample size (n=66).

Cal. set	Traits	No.	Range	Mean	SD	SEC	R^2_c	SECV	R^2_{cv}	RPD
a. Calibration of carotenoids on mashed samples.										
C16M66	TCC	63	10.09-25.81	17.72	3.67	0.42	0.99	1.23	0.89	2.98
	ATBC	59	4.91-16.42	11.40	3.17	0.25	0.99	0.82	0.93	3.87
b. Calibration of carotenoids on intact root samples										
C16I66	TCC	64	10.09-26.15	17.83	3.74	1.16	0.90	2.13	0.67	1.76
	ATBC	63	4.91-19.97	11.94	3.53	0.89	0.94	1.53	0.81	2.31

Validation of calibration models: Validation is very important in the development of a quantitative model using independent sets of samples different from the data employed in model construction (Pasquini 2003). Individual models developed from different calibration sets from mashed or intact root samples were used to predict the values of other sets in the same intact or mashed sample categories. As would be expected, especially where there were obvious differences in reference value protocols, the cross-prediction statistics based on coefficient of determination (R^2_p) were less than 50%

except in the case of using U16M for calibration and C16M for validation on DMC calibration (Table 2.7).

Table 2.7: Validation using different calibration sets on intact and mashed root samples for DMC

Calibration set	Validation set	SEP	SEP(C)	R^2
a. Cross-calibration set validations on intact root samples.				
U15I	U16I	133.37	7.45	0.03
U15I	C16I	147.18	6.07	0.04
U16I	U15I	65.48	17.82	0.28
U16I	C16I	7.91	3.51	0.39
C16I	U15I	28.65	19.20	0.18
C16I	U16I	12.50	7.29	0.19
Cross-calibration set validations on mashed root samples				
C16M	U16M	6.81	4.38	0.48
U16M	**C16M**	**3.10**	**2.59**	**0.72**

For independent validation of models, the mashed calibration set developed at CIAT was trimmed and divided into calibration and validation sets for the three traits – DMC, TCC and ATBC. Previously, the effect of trimming on the Vis/NIRS data was evaluated by comparing calibrations developed from untrimmed and trimmed sets. The result showed that there was no obvious variation or trend between the trimmed and untrimmed data sets (Appendix 2.3). Using the trimmed calibration and validation sets, models were built using the calibration set with larger number of samples and used to predict an independent validation set with fewer training sets (Scenario 1) and conversely, using the validation set to predict the values of the larger set (Scenario 2).

The average values from the two scenarios were used for independent calibration and validation of models for the three traits. The use of larger number of calibration (Scenario 1) was slightly higher for DMC and ATBC than TCC (Table 2.8). This probably highlights the role of calibration size on prediction accuracy. The coefficient of determination for prediction (R^2_p) ranged from 76% to 91%. On the average, R^2_p for ATBC was highest (91%) followed by TCC (88%) and DMC (80%). The same pattern was observed in RPD distribution. The standard error of prediction corrected for bias SEP(C) was lowest in ATBC (1.65 µg) and highest in TCC (2.36 µg) while DMC was 1.77 percent. The high R^2_p values (>80%) showed that the handheld Vis/NIRS device could be useful in quality and standardized phenotyping in cassava breeding especially for DMC, TCC and ATBC.

Table 2.8: Independent validation of models for DMC, TCC and ATBC

Trait	Cal	Val	SEP	SEP(C)	R^2_p	SD	RPD
DMC	Cal	Val	1.47	1.46	0.836	3.4	2.3
	Val	Cal	2.10	2.08	0.763	4.14	2.0
TCC	Cal	Val	2.64	2.52	0.859	6.62	2.6
	Val	Cal	2.23	2.19	0.901	6.28	2.9
ATBC	Cal	Val	1.70	1.59	0.908	5.21	3.3
	Val	Cal	1.70	1.71	0.902	4.8	2.8

Correlations of NIRS analyzed, specific gravity and oven-drying dry matter content (DMC) methods:

Compared to the current regression equation used by many breeding programs, DMC = 158.3SG – 142 (R^2 = 0.84) (Kawano, Fukuda, and Cenpukdee 1987; Fukuda et al. 2010; Pérez et al. 2011), the relationship between DMC and SG obtained from the NRCRI dataset was given as DMC = 67.33SG – 37.03 (R^2 = 0.23). The correlations among the four DMC methods showed positive relationships among the different methods (Table 2.9). The highest correlation (0.98) was between oven-drying method and NIRS-derived DMC on mashed root samples. The correlation between oven-drying method and NIRS-derived values from intact root was also very high (0.95) and similar to the relationship

between NIRS on intact and mashed root samples (0.94). There was a moderate correlation (0.49) between DMC by oven-drying and specific gravity methods.

Table 2.9: Correlations among the different DMC methods (n = 179).

	$NIRS_I$	$NIRS_M$	DM_V	DM_G
$NIRS_I$	1			
$NIRS_M$	0.94	1		
DM_V	0.95	0.98	1	
DM_G	0.54	0.49	0.49	1

$NIRS_I$ = DMC by portable NIRS on intact root samples; $NIRS_M$ = DMC by portable NIRS on mashed root samples; DM_V = DMC by oven method; DM_G = DMC by specific gravity method.

In addition, given that NIRS derived values ($NIRS_I$ and $NIRS_M$) were obtained from models trained with oven-dried reference DM values, we tried to compare values from both oven-dry and specific gravity references values. The calibration models were fitted with three passes as described earlier. Due to differential removal of outliers from the four calibration sets – calibration using SG on intact root ($SGNIRS_I$) and mashed samples ($SGNIR_M$) as well as from oven values on intact ($OvenNIRS_I$) and mashed samples ($OvenNIR_M$), only 159 common samples were for the correlation of the six datasets.

Comparing the calibration models from the different methods on both intact and mashed NIRS spectra, the standard errors of calibration (SEC) and cross-validations (SECV)

were all higher with SG methods than the oven models (Table 2.10). The R^2_c and R^2_{cv} from SG in both intact and mashed samples were less than 0.4 whereas, the R^2_c for oven methods was 0.91 on intact and 0.97 on mashed with R^2_{cv} of 0.66 (intact) and 0.84 (mashed) (Table 2.10).

Table 2.10: Comparison of DMC calibration from SG (intact and mashed) and Oven (intact and mashed) methods using 2016 NRCRI dataset.

Calibration Set	SEC	R^2_c	SECV	R^2_{cv}
SG_I	3.01	0.50	3.63	0.32
OV_I	1.38	0.91	2.76	0.66
SG_M	3.26	0.41	3.49	0.36
OV_M	0.88	0.97	1.87	0.84

While the correlations between DMC from oven method and NIRS-derived DMC from oven-NIRS model on intact and mashed samples remained the same (0.95 and 0.98, respectively), the removal of outliers during calibration, favored the correlation from SG – an increase from 0.49 to 0.64 (Table 2.11). The correlation between oven method and SG-NIRS derived values were 0.71 and 0.85 from intact and mashed samples, respectively (Table 2.11).

Table 2.11: Correlation of DMC values obtained from SG (intact and mashed) and Oven (intact and mashed) models with SG and Oven values using 2016 NRCRI dataset (n = 159).

	DM_{SG}	$SGNIRS_I$	$SGNIR_M$	DM_{Oven}	$OvenNIRS_I$	$OvenNIR_M$
DM_{SG}	1					
$SGNIRS_I$	0.72	1				
$SGNIR_M$	0.62	0.64	1			
DM_{Oven}	0.64	0.71	0.85	1		
$OvenNIRS_I$	0.67	0.79	0.83	0.95	1	
$OvenNIR_M$	0.64	0.7	0.88	0.98	0.94	1

DM_{SG} = DMC derived from SG; $SGNIRS_I$ = DMC derived from SG-NIRS model on intact roots; $SGNIR_M$ = DMC derived from SG-NIRS model on mashed roots; DM_{Oven} = DMC derived from Oven drying; $OvenNIRS_I$ = DMC derived from Oven-NIRS model on intact roots; $OvenNIR_M$ = DMC derived from Oven-NIRS model on mashed roots.

Although it is very important to standardize the drying conditions for the oven-drying method in different breeding programs, it might be necessary for each system to review the relationship between specific gravity and reference DMC by oven-drying and establish protocols for accurate sampling. The low R^2 value obtained in this study could be attributed to the sampling protocols, weight and number of roots used for specific gravity measurement (Fukuda et al. 2010; Pérez et al. 2011). Field-based specific gravity and for very large population is usually carried out before peeling, and cassava peels have been reported to constitute as high as 7.9% of the root size (Pérez et al. 2011)

and could even be higher with soil particles and fibrous neck still attached to the root. This could reduce the reported relation between DMC derived by specific gravity and oven-drying, which in most cases was carried out after peeling (Pérez et al. 2011). On the other hand, the use of Vis/NIRS, could help to address the challenges associated with the existing methods while improving the overall quality of phenotyping in cassava.

Conclusion

From the results of this study, the choice of mathematical pre-processing is a very important step in developing a robust calibration model and the choice of pre-treatment method might be influenced by sampling methods. Calibration models developed with mashed samples were clearly better than intact root samples although the calibration performance for some of the intact root models were still adequate for screening purposes. Also, since the correlation between DMC analysis on intact and mashed root samples was very high, the Vis/NIRS could be employed for initial screening in the field before further extensive laboratory analyses. However, with improved spectra collection protocols and increasing the number of scanning points per root, we hope to further improve calibration performance from intact root samples given that mashing requires additional resources including time and cost of harvesting. The handheld Vis/NIRS has great potential for standardized and unbiased analyses of traits in cassava breeding. It provides a good alternative for the evaluation and improvement of many novel traits which have been difficult or costly to measure before now. In addition to being a non-destructive analytical tool that only requires minimal sample preparation, the portable NIRS is very useful in direct field analyses and will help reduce sample degradation. When compared to the conventional laboratory methods for DMC and carotenoids in cassava breeding, NIRS technique is rapid and cost-effective. It is a good alternative to quality and unbiased evaluation of traits especially in low-cost breeding programs.

Acknowledgements

We acknowledge the efforts of the NextGen Cassava team at NRCRI, Umudike, Nigeria (Ahamefule Nwogu, Kelechi Njoku, Ikechukwu Nnaji, Chinedozi Amaefula, Benjamin Ochulorugo, Ivory Ndukwe, and Precious Udoka), the breeding program and HarvestPlus team at CIAT, Cali, Columbia especially John Belalcazar, Luis Londono, Angelica Jaramillo and Talsma Elise for their contributions in data collection and laboratory assistance. Uche Godfrey Okeke participated in the discussion and the choice of the portable Vis/NIRS.

APPENDICES

Appendix 2.1: Calibration for carotenoids from mashed samples using the entire calibration set from CIAT.

Cal. set	Traits (µg)	No.	Range	Mean	SD	SEC	R^2_c	SECV	R^2_{cv}	RPD
C16M	VIO	164	0.09-0.92	0.47	0.17	0.042	0.94	0.11	0.61	1.55
	LUT	100	0.02-1.27	0.36	0.32	0.11	0.88	0.24	0.42	1.33
	15CBC	163	0.01-0.44	0.23	0.10	0.01	0.99	0.04	0.83	2.50
	13CBC	165	0.04-2.46	1.22	0.59	0.06	0.99	0.28	0.78	2.11
	AC	74	0.03-0.10	0.07	0.02	0.01	0.80	0.01	0.65	2.00
	9CBC	171	0.10-2.51	0.98	0.51	0.06	0.99	0.24	0.77	2.13
	PHY	87	0.96-13.79	5.85	2.71	0.80	0.91	2.07	0.41	1.31

Appendix 2.2: Calibrations for additional carotenoids from mashed (a) and intact (b) root samples using common samples (n=66)

Cal. set	Traits	No.	Range	Mean	SD	SEC	R^2_c	SECV	R^2_{cv}	RPD
a. Calibration of carotenoids on mashed samples										
C16M66	VIO	62	0.22-0.84	0.52	0.14	0.02	0.98	0.10	0.47	1.40
	LUT	51	0.04-2.36	0.65	0.54	0.20	0.87	0.44	0.33	1.23
	15CBC	64	0.18-0.44	0.29	0.06	0.02	0.91	0.04	0.60	1.50
	13CBC	63	0.72-2.42	1.57	0.40	0.08	0.96	0.22	0.68	1.82
	AC	35	0.04-0.10	0.07	0.02	0.008	0.73	0.01	0.47	2.00
	9CBC	63	0.44-2.04	1.22	0.36	0.07	0.96	0.21	0.65	1.71
	PHY	63	0.96-13.79	5.86	3.01	0.77	0.93	2.39	0.36	1.26
b. Calibration of carotenoids on intact root samples										
C16M66	VIO	60	0.22-0.84	0.52	0.14	0.07	0.71	0.11	0.38	1.27
	LUT	46	0.04-1.32	0.51	0.34	0.30	0.22	0.32	0.10	1.06
	15CBC	62	0.18-0.44	0.29	0.06	0.02	0.88	0.04	0.44	1.50
	13CBC	64	0.72-3.05	1.59	0.43	0.22	0.73	0.37	0.27	1.16
	AC	34	0.04-0.10	0.08	0.01	0.01	0.67	0.01	0.46	1.00
	9CBC	64	0.44-2.04	1.22	0.36	0.15	0.82	0.29	0.34	1.24

	PHY	61	0.96-13.79	5.69	2.91	0.40	0.98	1.97	0.53	1.48

Appendix 2.3: Calibrations for DMC, TCC and ATBC using trimmed and untrimmed ASD spectra

Traits	Sets	Status	SEC	R^2_c	SECV	R^2_{cv}	SD	RPD
DMC	Cal.	Untrimmed	0.49	0.987	0.93	0.953	4.34	4.7
		Trimmed	0.63	0.979	0.91	0.956	4.36	4.8
	Val.	Untrimmed	0.37	0.990	0.93	0.937	3.75	4.0
		Trimmed	0.33	0.993	0.87	0.946	3.79	4.4
TCC	Cal.	Untrimmed	0.58	0.994	1.94	0.928	7.24	3.7
		Trimmed	0.51	0.995	1.78	0.937	7.14	4.0
	Val.	Untrimmed	0.90	0.985	2.28	0.897	7.20	3.2
		Trimmed	1.15	0.972	2.41	0.875	6.89	2.9
ATBC	Cal.	Untrimmed	0.54	0.990	1.31	0.943	5.48	4.2
		Trimmed	0.50	0.991	1.35	0.937	5.40	4.0
	Val.	Untrimmed	0.64	0.987	1.68	0.905	5.51	3.3
		Trimmed	1.21	0.955	2.09	0.865	5.74	2.7

REFERENCES

Arganosa, G. C., Warkentin, T. D., Racz, V. J., Blade, S., Phillips, C., & Hsu, H. (2006). Prediction of crude protein content in field peas using near infrared reflectance spectroscopy. *Canadian Journal of Plant Science*, *86*, 157–159. Retrieved from http://www.nrcresearchpress.com/doi/pdf/10.4141/P04-195

ASD. (2008). ViewSpec Pro ™ User Manual. *ASD Document 600555 Rev. A*. Retrieved from http://support.asdi.com/Document/Viewer.aspx?id=31

Asrat, S., Yesuf, M., Carlsson, F., & Wale, E. (2010). Farmers' preferences for crop variety traits: Lessons for on-farm conservation and technology adoption. *Ecological Economics*, *69*(12), 2394–2401. http://doi.org/10.1016/j.ecolecon.2010.07.006

Barnes, R. J., Dhanoa, M. S., & Lister, S. J. (1989). Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra. *Applied Spectroscopy*, *43*(5), 772–777. http://doi.org/10.1366/0003702894202201

Belalcazar, J., Dufour, D., Andersson, M. S., Pizarro, M., Luna, J., Londoño, L., … Ceballos, H. (2016). High-throughput phenotyping and improvements in breeding cassava for increased carotenoids in the roots. *Crop Science*, *56*(6), 2916–2925. http://doi.org/10.2135/cropsci2015.11.0701

Blanco, M., Coello, J., Iturriaga, H., Maspoch, S., & De La Pezuela, C. (1997). Effect of data preprocessing methods in near-infrared diffuse reflectance spectroscopy for the determination of the active compound in a pharmaceutical preparation. *Applied Spectroscopy*, *51*(2), 240–246. http://doi.org/10.1366/0003702971939947

Blanco, M., & Villarroya, I. (2002). NIR spectroscopy: a rapid-response analytical tool. *TrAC Trends in Analytical Chemistry*, *21*(4), 240–250. http://doi.org/http://dx.doi.org/10.1016/S0165-9936(02)00404-1

Büning-Pfaue, H. (2003). Analysis of water in food by near infrared spectroscopy. *Food Chemistry*, *82*(1), 107–115. http://doi.org/10.1016/S0308-8146(02)00583-6

Cabrera-Bosquet, L., Crossa, J., von Zitzewitz, J., Serret, M. D., & Luis Araus, J. (2012). High-throughput Phenotyping and Genomic Selection: The Frontiers of Crop Breeding Converge. *Journal of Integrative Plant Biology*, *54*(5), 312–320. http://doi.org/10.1111/j.1744-7909.2012.01116.x

Campbell, M. R., Mannis, S. R., Port, H. A., Zimmerman, A. M., & Glover, D. V. (1999). Prediction of starch amylose content versus total grain amylose content in corn by near-infrared transmittance spectroscopy. *Cereal Chemistry*, *76*(4), 552–557. http://doi.org/10.1094/CCHEM.1999.76.4.552

Ceballos, H., Kawuki, R. S., Gracen, V. E., Yencho, G. C., & Hershey, C. H. (2015). Conventional breeding, marker-assisted selection, genomic selection and inbreeding in clonally propagated crops: a case study for cassava. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *128*(9), 1647–67. http://doi.org/10.1007/s00122-015-2555-4

Ceballos, H., Morante, N., Sánchez, T., Ortiz, D., Aragón, I., Chávez, A. L., … Dufour, D. (2013). Rapid cycling recurrent selection for increased carotenoids content in cassava roots. *Crop Science*, *53*(6), 2342–2351.

http://doi.org/10.2135/cropsci2013.02.0123
D'Acqui, L. P., Pucci, A., & Janik, L. J. (2010). Soil properties prediction of western Mediterranean islands with similar climatic environments by means of mid-infrared diffuse reflectance spectroscopy. *European Journal of Soil Science*, *61*(6), 865–876. http://doi.org/10.1111/j.1365-2389.2010.01301.x
Davrieux, F., Dufour, D., Dardenne, P., Belalcazar, J., Pizarro, M., Luna, J., … Jaramillo, A. (2016). LOCAL regression algorithm improves near infrared spectroscopy predictions when the target constituent evolves in breeding populations. *Journal of Near Infrared Spectroscopy*, *24*(2), 109–117. http://doi.org/10.1255/jnirs.1213
De Alencar Figueiredo, L. F., Davrieux, F., Fliedel, G., Rami, J. F., Chantereau, J., Deu, M., … Mestres, C. (2006). Development of NIRS equations for food grain quality traits through exploitation of a core collection of cultivated sorghum. *Journal of Agricultural and Food Chemistry*, *54*(22), 8501–8509. http://doi.org/10.1021/jf061054g
dos Santos, C. A. T., Lopo, M., Páscoa, R. N. M. J., & Lopes, J. A. (2013). A Review on the Applications of Portable Near-Infrared Spectrometers in the Agro-Food Industry. *Applied Spectroscopy*, *67*(11), 1215–1233. http://doi.org/10.1366/13-07228
Fox, G. P., O'Donnell, N. H., Stewart, P. N., & Gleadow, R. M. (2012). Estimating hydrogen cyanide in forage sorghum (Sorghum bicolor) by near-infrared spectroscopy. *Journal of Agricultural and Food Chemistry*, *60*(24), 6183–6187. http://doi.org/10.1021/jf205030b
Freschet, G. T., Barthès, B. G., Brunet, D., Hien, E., & Masse, D. (2011). Use of Near Infrared Reflectance Spectroscopy (NIRS) for Predicting Soil Fertility and Historical Management. *Communications in Soil Science and Plant Analysis*, *42*(14), 1692–1705. http://doi.org/10.1080/00103624.2011.584597
Fukuda, W. M. G., Guevara, C. L., Kawuki, R., & Ferguson, M. E. (2010). Selected morphological and agronomic descriptors for the characterization of cassava. *Internation Institute of Tropical Agriculture*, 19.
Geladi, P., MacDougall, D., & Martens, H. (1985). Linearization and scatter-correction for NIR reflectance spectra of meat. *Applied Spectroscopy*, *39*(3), 491–500.
Jannink, J.-L., Lorenz, A. J., & Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. *Briefings in Functional Genomics*, *9*(2), 166–177. http://doi.org/10.1093/bfgp/elq001
Kawano, K., Fukuda, W. M. G., & Cenpukdee, U. (1987). Genetic and Environmental Effects on Dry Matter Content of Cassava Root1. *Crop Science*, *27*(1), 69. http://doi.org/10.2135/cropsci1987.0011183X002700010018x
Li, D., Liu, Y., Chen, Y., Wang, X., & Zhou, G. (2011). Study on Pretreatment Algorithm of Near Infrared Spectroscopy, 623–632. Retrieved from http://download.springer.com/static/pdf/977/chp%253A10.1007%252F978-3-642-18336-2_76.pdf?originUrl=http%3A%2F%2Flink.springer.com%2Fchapter%2F10.1007%2F978-3-642-18336-

2_76&token2=exp=1495487737~acl=%2Fstatic%2Fpdf%2F977%2Fchp%25253A10.1007%25252F978-3-64

Lopez, A., Arazuri, S., Garcia, I., Mangado, J., Jaren, C., & Accepted, J. (2013). Review A REVIEW ON THE APPLICATION OF NEAR-INFRARED SPECTROSCOPY FOR THE ANALYSIS OF POTATOES FOR THE ANALYSIS OF POTATOES. http://doi.org/10.1021/jf401292j

Lu, G., Huang, H., & Zhang, D.-P. (2006). Application of near-infrared spectroscopy to predict sweetpotato starch thermal properties and noodle quality. *Journal of Zhejiang University. Science. B*, *7*(6), 475–81. http://doi.org/10.1631/jzus.2006.B0475

Lu, G., Huang, H., & Zhang, D. (2006). Prediction of sweetpotato starch physiochemical quality and pasting properties using near-infrared reflectance spectroscopy. *Food Chemistry*, *94*(4), 632–639. http://doi.org/10.1016/j.foodchem.2005.02.006

Manley, M. (2014). Near-infrared spectroscopy and hyperspectral imaging: non-destructive analysis of biological materials. *Chem. Soc. Rev.*, *43*(24), 8200–8214. http://doi.org/10.1039/C4CS00062E

Marten, G., Shenk, J., & Barton, F. (1989). Near infrared reflectance spectroscopy (NIRS): analysis of forage quality. *U.S. Department of Agriculture, Agriculture Handbook*, *643*, 1–110.

Naes, T., Isaksson, T., Fearn, T., & Davies, T. (2002). A User-friendly Guide to Multivariate Calibration and Classification. *NIR Publications*, *46*(1), 7–289. http://doi.org/10.1198/004017004000000167

Ortiz, D., Sánchez, T., & Morante, N. (2011). Sampling strategies for proper quantification of carotenoid content in cassava breeding. *Plant Breed. Crop …*. Retrieved from http://r4d.dfid.gov.uk/pdf/outputs/misc_crop/ortiz-et-al.pdf

Pasquini, C. (2003). Near Infrared Spectroscopy: fundamentals, practical aspects and analytical applications. *Journal of the Brazilian Chemical Society*, *14*(2), 198–219. http://doi.org/10.1590/S0103-50532003000200006

Pérez, J. C., Lenis, J. I., Calle, F., Morante, N., Sánchez, T., Debouck, D., & Ceballos, H. (2011). Genetic variability of root peel thickness and its influence in extractable starch from cassava (Manihot esculenta Crantz) roots. *Plant Breeding*, *130*(6), 688–693. http://doi.org/10.1111/j.1439-0523.2011.01873.x

Pizarro, C., Esteban-Díez, I., Nistal, A. J., & González-Sáiz, J. M. (2004). Influence of data pre-processing on the quantitative determination of the ash content and lipids in roasted coffee by near infrared spectroscopy. *Analytica Chimica Acta*, *509*(2), 217–227. http://doi.org/10.1016/j.aca.2003.11.008

Rinnan, Å., Berg, F. van den, & Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *TrAC - Trends in Analytical Chemistry*. http://doi.org/10.1016/j.trac.2009.07.007

Rodriguez-Amaya, D. ., & Kimura, M. (2004). HarvestPlus Handbook for Carotenoid Analysis. *HarvestPlus Technical Monographs*, 59. Retrieved from http://ebrary.ifpri.org/utils/getfile/collection/p15738coll2/id/125148/filename/125149.pdf

Roggo, Y., Chalus, P., Maurer, L., Lema-Martinez, C., Edmond, A., & Jent, N.

(2007). A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. *Journal of Pharmaceutical and Biomedical Analysis*, *44*(3 SPEC. ISS.), 683–700. http://doi.org/10.1016/j.jpba.2007.03.023

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

Sánchez, T., Chávez, A. L., Ceballos, H., Rodriguez-Amaya, D. B., Nestel, P., & Ishitani, M. (2006). Reduction or delay of post-harvest physiological deterioration in cassava roots with higher carotenoid content. *Journal of the Science of Food and Agriculture*, *86*(4), 634–639. http://doi.org/10.1002/jsfa.2371

Shenk, J. S., & Westerhaus, M. O. (1991). Population Definition, Sample Selection, and Calibration Procedures for Near Infrared Reflectance Spectroscopy. *Crop Science*, *31*(2), 469. http://doi.org/10.2135/cropsci1991.0011183X003100020049x

Stevens, A., & Ramirez Lopez, L. (2013). An introduction to the prospectr package, 1–22. Retrieved from https://cran.r-project.org/web/packages/prospectr/vignettes/prospectr-intro.pdf

Stuart, B. H. (2004). *Infrared Spectroscopy: Fundamentals and Applications*. *Methods* (Vol. 8). http://doi.org/10.1002/0470011149

Terhoeven-Urselmans, T. (2007). *Usefulness of near infrared spectroscopy to assess the composition and properties of soil, litter and growing media*. Kassel Univ. Press.

Tillmann, P., Reinhardt, T.-C., & Paul, C. (2000). Networking of near infrared spectroscopy instruments for rapeseed analysis: a comparison of different procedures. *J. Near Infrared Spectrosc*, *8*, 103–107. Retrieved from http://journals.sagepub.com/doi/pdf/10.1255/jnirs.269

Wang, Z., Kawamura, K., Sakuno, Y., Fan, X., Gong, Z., & Lim, J. (2017). Retrieval of Chlorophyll-a and Total Suspended Solids Using Iterative Stepwise Elimination Partial Least Squares (ISE-PLS) Regression Based on Field Hyperspectral Measurements in Irrigation Ponds in Higashihiroshima, Japan. *Remote Sensing*, *9*(3), 264. http://doi.org/10.3390/rs9030264

Williams, P., & Sobering, D. (1993). Comparison of commercial near infrared transmittance and reflectance instruments for analysis of whole grains and seeds. *Journal of Near Infrared Spectroscopy*, *1*(1), 25–32. http://doi.org/10.1255/jnirs.3

CHAPTER 3: NON-LINEAR CALIBRATION OF PORTABLE AND NEAR INFRARED SPECTROMETER (VIS/NIRS) IN TRAIT CORRELATIONS, GENOME-WIDE ASSOCIATION STUDIES AND GENOMIC PREDICTIONS OF CAROTENOIDS CONCENTRATION IN CASSAVA ROOTS.

Abstract:

For quantitative and high throughput screening of large numbers of cassava samples of carotenoid concentration in cassava roots, a non-linear calibration model – random forest (RF) was developed. The model was later used to analyze spectral data from the training population (TP) of NRCRI, Umudike Nigeria for total carotenoid content (TCC) and other individual carotenoids. Trait correlations (phenotypic and genetic), genome-wide association studies (GWAS) and genomic predictions using linear and non-linear, single and multiple traits as well as between one and two-stage models were compared. The calibration performance for most of the carotenoids were high based on the R^2 in calibration (0.59 to 0.92, except in phytoene) and correlations between reference and predicted values using independent samples (0.62 to 0.97). Very high and positive phenotypic and genotypic correlations (>0.75 except in violaxanthin and phytoene) were obtained between TCC and most of the PVA carotenoids (PVAC) while other forms of association existed among the carotenoid components. The GWAS confirmed a previously identified region on chromosome 1 as well as other regions that are associated with TCC variation and the individual carotenoids. Overall, the use of non-linear prediction model showed improved accuracies compared to linear GS model for predicting GEBVs of cassava clones. The multiple traits model was relatively higher than the single traits linear models in two-stage but not in one-stage approaches. For most of the traits, the one-stage GS prediction accuracies, although computationally intensive, was higher than the two-stage approach. This study is one of the initial attempts in dissecting and understanding the genomics of TCC and its individual components in cassava. It demonstrates the efficiency of incorporating Vis/NIRS with modern breeding tools for a large-scale evaluation and breeding for improved carotenoids content in cassava.

Keywords: Cassava; GS; GWAS; Calibration; near infra-red Spectroscopy (NIRS), Random forest; Single trait GS; Multi-trait GS; Non-linear GS models; Carotenoids.

Introduction:

Carotenoids are a class of more than 750 naturally occurring pigments synthesized by plants, algae, and photosynthetic bacteria (Armstrong and Hearst 1996; Owens et al. 2014). They are lipophilic substances absorbed from the small intestine along with other lipids and reappearing in the lipoprotein fractions of the plasma, as well as in erythrocytes and leucocytes (Strobel, Tinz, and Biesalski 2007). Structurally, they

contain C_{40} tetraterpenoids formed from eight C_5 isoprenoid units with conjugated double bonds where the carbon units are bonded together by alternating single and double bonds and the symmetry is imparted by a central molecule inversion (Rodriguez-Amaya and Kimura 2004). Their light absorption is influenced by the amount of conjugated double bonds present in the carotenoid and ranges in the 400 -500 nm region of the visible spectrum. As result, carotenoids vary in color from red, orange, and yellow (Hammond and Renzi 2013; Tosato et al. 2016). Also, changes in geometrical configuration about the double bonds result in the existence of many cis- and trans-isomers and hydroxylated, oxidized, hydrogenated or ring-containing derivatives (Rodriguez-Amaya and Kimura 2004). Carotenes (e.g. β-carotene) are classified as hydrocarbon carotenoids while xanthophylls (e.g. Lutein) contain oxygen (Hammond and Renzi 2013; D. B. D. . Rodriguez-Amaya and Kimura 2004).

Carotenoids are well known for their nutritional and health benefits in the prevention of a variety of major human diseases, including certain cancers and eye diseases (Krinsky and Johnson 2005; Paiva and Russell 2013). Of uttermost importance is the vitamin A activity of PVAC including beta-carotene, alpha-carotene, beta-cryptoxanthin and gamma-carotene, which are generally characterized by unsubstituted beta-ionone rings. Vitamin A is essential for growth and differentiation of a number of cells and tissues. It plays an important role in the healthy development of the fetus and the newborn (Strobel, Tinz, and Biesalski 2007). The lack of adequate intake of vitamin A has been associated with impaired vision, poor immunity, retarded growth and even death, more especially among children and pregnant or nursing mothers (Ceballos et al. 2013; Strobel, Tinz, and Biesalski 2007; Bechoff et al. 2015).

Cassava plays an important role in the diets of many people, majorly in the Sub-Saharan Africa where over 600 million people depend on it to meet their energy requirement (Oliveira et al. 2014; Rabbi et al. 2013). Although, rated as the fourth most important basic food after rice, wheat, and maize worldwide (Nweke 2004; Ceballos et al. 2017), the nutritional quality of cassava roots in general is low, and contains mainly carbohydrates (Hernán Ceballos et al. 2017). In order to curb the impact of the low nutritional quality, there is an ongoing effort to overcome especially the vitamin A deficiency (VAD) of many staple foods including cassava, maize, potatoes, etc., taking advantage of existing genetic variability among the different crops (Ceballos et al. 2017; Ceballos et al. 2013; Mugode et al. 2014). The targeted increase in micro-nutrient content of cassava including PVAC is desirable to alleviate the VAD problem, especially for those within the poverty bracket who cannot afford healthy and balanced nutrition from other more expensive food sources. The bio-fortification initiative in cassava has already led to a substantial increase in the proportion of carotenoids in cassava roots in many breeding programs (Belalcazar et al 2016; Ceballos et al. 2013). The recorded success among other things was possible because of the advancement and application of new analytical tools (Ceballos, et al. 2016; Pérez et al. 2011). Usually, the use of color intensity could be highly challenging and limited to qualitative classification of clones into white, cream and yellow categories (Ceballos et al. 2017). Alternatively, the use of high-performance liquid chromatography (HPLC) or UV-Visible spectrophotometer are low-throughput and require skilled labor, constant chemical reagents especially for HPLC, as well as favorable laboratory conditions to operate (Ceballos et al. 2013; Belalcazar et al 2016). However, such laboratory facilities

and conditions are lacking in low resource breeding programs and out-stations where multi-location evaluations take place. Hence, the need for more analytical tools that will facilitate quantitative and high-throughput assessment of both known and novel traits including micro-nutrients. Recently, the calibration and use of visible and near infra-red spectroscopy (Vis/NIRS) has been demonstrated to promote high-throughput and enables the quantification of individual carotenoids (Sánchez et al. 2014; Belalcazar et al 2016), a development that will not only improve phenotyping for these traits but equally increases the understanding of the underlying genetics as well as guide the improvement of PVAC in cassava.

Although, linear regression models have been useful for NIRS calibration with simple and easy to interpret results, they are often limited by nonlinear effects including baseline drift, light scattering effect, multicollinearity, etc. on spectra. Recently, non-linear models, especially random forest (RF) has been used to model both linear and nonlinear multivariate calibration and recommended for spectra analyses due to its comparable accuracy, mathematical simplicity, computational efficiency and robustness to noise (Ghasemi and Tavakoli 2013; Lee et al. 2012; Breiman 2001). Compared to most of the popular linear as well as other non-linear calibration models, the principal component regression (PCR) and partial least square regression (PLSR) algorithms perform linear regression on the factor analysis components which arguably lack any physical meaning, and some other nonlinear models are either inefficient in modeling high-dimensional or not robust enough in handling noisy data sets (Ghasemi and Tavakoli 2013; Cristianini and Shawe-Taylor 2000; Wold, Sjöström, and Eriksson 2001; Andersson 2009). On the other hand, RF has been effective in multivariate

calibrations from modern measuring instruments including spectrometers, chromatographs and sensor batteries where it has been used to provide more interpretable algorithm and offer adequate fine-tuning mechanism to control over-fitting, and adequately deal with collinearity, associated with most spectroscopic data (Ghasemi and Tavakoli 2013; Svetnik et al. 2003; Sila, Shepherd, and Pokhariyal 2016). The lack of adequate phenotyping tools especially in dissecting total carotenoid content (TCC) into its individual components could have restricted the genetic analyses aimed at understanding the key genes involved in natural variation for PVAC in cassava mainly to TCC. Genome-wide association studies (GWAS), which leverages on the available marker information distributed throughout the genome, has been useful in unraveling the genomic region associated with carotenoids variation in cassava (I. Y. Rabbi et al. 2017; Esuma et al. 2016). The limited information on the relative genetic control and relationships (genotypic and phenotypic correlations) among the individual carotenoid components is still a challenge in breeding for these traits in cassava. Naturally, carotenoids are predominately present in all-trans configuration while the other forms of isomerization such as cis isomers are present and are more polar, less prone to crystallization, more soluble in oils and hydrocarbon solvents (Paiva and Russell 2013; Castenmiller and West 1998). Many factors including processing, species of carotenoid, molecular linkage, amount of carotenoids consumed in a meal, matrix in which the carotenoid is incorporated, effectors of absorption and bioconversion, nutrient status of the host, genetic factors, host-related factors and mathematical interactions have been recognized to play important roles in the retention, bioavailability and bioconversion of carotenoids (Bechoff et al. 2015; Mugode et al. 2014; Castenmiller

and West 1998; Strobel, Tinz, and Biesalski 2007). The effect of the matrix in which the carotenoids is incorporated is very important, and many interactive responses have been reported including a positive bioavailability and bioconversion interaction between β-carotene and concentrations of α-carotene as well as negative interactions between β-carotene and lutein, lycopene, and canthaxanthin (Castenmiller and West 1998). Lutein was reported to interfere with the conversion of β-carotene to retinol and may explain in part, the low conversion to retinol of β-carotene from dark- green leafy vegetables (van Vliet et al. 1996).

β-carotene has been associated with the highest vitamin A activity on molar basis, however, the vitamin A activities of other PVAC and their relative proportion to β-carotene (for example, cis isomers - 50%, α-carotene - 29%, β-cryptoxanthin - 55% etc.) has been identified (FAO/WHO 1998; van Vliet et al. 1996). The metabolism of all-trans- and 9-cis-β-carotene have shown to produce similar amounts of retinoic acid, with 9-cis-β-carotene giving rise to equal amounts of all-trans- and 9-cis-retinoic acid (Hebuterne et al. 1995; Castenmiller and West 1998). While it is important to fully understand the bioavailability of carotenoids from cassava, there is need for adequate understanding of both the genetic variation existing within different breeding programs and the underlying genotypic relationships among the various carotenoids components in the current effort in improving this group of traits.

Unlike GWAS, genomic selection (GS) is a breeding technology that is used to predict the genetic potential of individuals in a breeding program without necessarily uncovering the underlying genes and QTL behind the traits of interest (Iwata et al. 2013; Jannink, Lorenz, and Iwata 2010). It has potential to accelerate genetic gain over time,

shorten breeding cycles and cost in breeding programs (Ben J. Hayes et al. 2010; Eder Jorge de Oliveira et al. 2012; Habier, Fernando, and Dekkers 2009; Wolfe et al. 2017). As the field continues to grow and new computational methods are being developed, non-linear GS models have been shown to be useful in estimating total genetic values other than just breeding values of individuals (Heslot et al. 2012; Lorenz et al. 2011; Wolfe et al. 2017). Also, although often used in the analyses of multi-environment trials for its simplicity and computational efficiency, the two-stage approach involving the computation of adjusted means for genotypes/clones followed by prediction of genomic breeding values in the second stage has been described as a good approximation of the single-stage approach which models the entire observed data at the level of individual plots and fully account for the entire variance–covariance structure of the observed data (Schulz-Streeck, Ogutu, and Piepho 2013; Piepho et al. 2012).

In this study, we used a nonlinear model in developing calibration models for some carotenoids in cassava and employed the models in analyzing the training population of a national breeding program - National Root Crops Research Institute (NRCRI), Umudike in Nigeria. We examined the correlations – phenotypic and genotypic, and the underlying genomic regions associated with the different carotenoid components and demonstrated the potential of using GS for the rapid improvement of these traits. We compared the conventional predictions of breeding values with a RF model that has the potential to capture non-additive signals. While many GS predictions are usually performed on a single trait basis, the use of multi-trait models have shown improvements in predictions of multiple traits, taking advantage of their correlations (Fernandes et al. 2017; Jia and Jannink 2012; Okeke et al. 2017). Therefore, in addition,

we compared predictions of single to multiple traits GS models for the improvement of carotenoids concentration in cassava roots.

Materials and methods:

Training population and spectra collection:

NRCRI has a training population that is being used for the implementation of GS in cassava (Wolfe et al. 2017). The germplasm consists of two different trials – training population 1 (TP1) and Training population 2 (TP2) which are based on the initial and subsequent germplasm collection for GS implementation. These two trials were further separated into sets for easy management and data control. In this study, TP1 was evaluated in one location at Umudike whereas TP2 was evaluated in three locations including Umudike, Otobi and Kano experimental stations in 2015/2016 cropping seasons. The trials were established as randomized incomplete block with three replications per location and plot size of five plants. A total of 594 clones from the two trials were used for this study. The origin of the NRCRI clones has been described and most of the breeding materials were developed from various forms of recombination with germplasm introduced from the International Center for Tropical Agriculture (CIAT), Cali-Palmira, Colombia and some other breeding materials shared between NRCRI and the International Institute for Tropical Agriculture (IITA), Ibadan, Nigeria (Wolfe et al. 2016).

Spectral data were collected using a portable Vis/NIRS (QualitySpec Trek: S-10016, ASD Inc.) from mashed root samples. Between two to three sizeable roots (arbitrary) were selected per plot for evaluation. The selected roots were peeled with knives, washed and homogenized into a paste-like mash using a portable power-operated grater.

Spectral data were collected from homogenized mashed samples in quartz sampling cups placed against the window of the portable Vis/NIRS device. Each sample was collected in two replications per clone. The internal number of spectra mean for each final spectra output was set to fifty scans per spectrum, which means that each spectrum is a mean of fifty internal iterations.

Calibration and analyses of carotenoids from NRCRI training population:

The calibration set used for model development with the portable Vis/NIRS device (350 – 2500nm wavelength in 1nm range) on mashed cassava samples has been described (Ikeogu et al. 2017). A total of 173 samples with both reference wet chemistry and spectra values were processed at CIAT in 2016. In order to overcome the challenges of converting the Vis/NIRS spectra data to conform to the customized software format using a commercial software, the VIS/NIRS calibration was performed in R platform, a free software environment for statistical computation and graphics (R Core Team 2017). The development of calibration in R besides saving cost, is necessary in promoting the ease of reproducing the calibration process.

Using TCC, we first assessed the calibration performance of two linear – PCR and PLSR and nonlinear RF calibration models.

***Principal component regression (PCR)*:** PCR uses the principal components provided by PCA to perform regression on the sample property to be predicted (Metrohm 2013; Chen and Wang 2001). PCA suppresses the spectral collinearity although there is no guarantee that the computed principal components are correlated to the studied property (Roggo et al. 2007; Metrohm 2013).

Partial least squares regression (PLSR): PLSR finds the directions of greatest variability by considering both spectral and target-property information (Tobias 1995; Roggo et al. 2007). The goal of the PLSR is to establish a linear link between spectral data and the reference values. It models both the spectral and reference values in order to find out the variables in the spectral matrix that will best describe the reference vector while reducing the dimensionality of the regression problem by using the minimum numbers of latent values (Ghasemi and Tavakoli 2013). This can be explained by the representation of the spectra in the space of wavelengths in order to show directions that will be linear combinations of wavelengths called factors which best describe the studied property (Roggo et al. 2007).

Random Forest (RF):

Random Forest is a predictor consisting of a collection of randomized base regression trees fused to form an aggregated regression estimate (Breiman 2001; Biau 2010). The process includes the random selection of samples with replacement from a given data set (calibration set) in order to create different trees (bootstrap sampling). For each tree in the bootstrapped set, a modified unpruned classification and regression tree (CART) algorithm is used to split at each node instead of testing the performance of all the variables (Breiman 2001; Biau 2010). It evaluates the performance of different number of randomly selected variables (mtry). Each tree grows until it reaches a predefined minimum number of nodes (nodesize). Following the concept of consensus modeling, the average of the prediction values of all trees is the final predicted value for that sample. RF is considered as one of the most accurate general-purpose learning

techniques available and has been reported to be fast and easy to implement with the capacity to handle a very large number of input variables without over-fitting (Biau 2010; Ghasemi and Tavakoli 2013).

Standard Normal Variate and De-trending (SNVD) spectra pre-treatment on (D = 2, G = 5, S1 = 2, S2 = 1) mathematical treatments, where D indicates the derivative order number (0 indicates no derivation, 1 means the first derivative, and so on), G indicates the gap (the number of data points over which derivation is computed), S1 indicates the number of data points in the first smoothing (1 means no smoothing) and S2 indicates the number of data points in the second smoothing (1 means no smoothing) was adopted to correct for external interferences on the spectral data. For validation, the total calibration set (n=173) was divided into training and testing set in a ratio of 3:1, and a repeated cross-validation was used for internal cross-validation within the training set. The calibration was iterated 10 times and 500 trees (ntree) were used in the initial case of RF. The reported calibration statistics include: correlation between the predicted and actual values of the training set (r_c), correlation between the predicted values using the model developed from the training set and actual values of the test set (r_{cv}), the coefficient of determination in calibration (R_c^2) and root mean squared error (RMSE) of the training model from the three regression methods.

Over 4000 spectral data from the NRCRI TP (TP1 and TP2) in the three locations – Umudike, Otobi and Kano of NRCRI were analyzed using the generated RF calibration model for all the measured carotenoids including TCC, all-trans β-carotene (ATBC), violaxanthin (VIO), Lutein (LUT), 15-Cis beta-carotene (15CBC), 13-Cis beta-carotene (13CBC), Alpha carotene (AC), 9-Cis beta-carotene (9CBC) and phytoene (PHY).

Genotype Data:

The genotype data used in this study have been previously described (Wolfe et al. 2016; Wolfe et al. 2017). The data were generated using Genotyping by sequencing (GBS) technique with ApeK1 restriction enzyme at read lengths of 100 bp. SNP calls were carried out with the TASSEL GBS pipeline V4 (Glaubitz et al., 2014) and aligned to the cassava reference genome (Goodstein et al. 2012; Bredeson et al. 2016). Individuals with more than 80% missing SNP calls and markers with more than 60% missing were removed. Marker data were converted to a dosage format and missing data were imputed with Beagle (version 4.0) (Browning and Browning 2008). After filtering based on MAF > 0.01, a total of 114884 SNP markers were used for genomic predictions.

GS models: Single trait (ST) – one vs two stage linear and non-linear models:

A two-stage single trait GS approach was used to estimate genomic breeding values for each of the carotenoids. Estimated genetic values (EGVs) were first derived using raw phenotype data corrected for location, trial, planting sets and replications. The EGVs were obtained as the best linear unbiased predictions (BLUPS) extracted from a linear mixed model that was fitted with the *lmer* package in R (Bates et al. 2014; R Core Team 2017). The fitted model was given as:

$$y_{ijk} = \mu + Xloc + Z_{clone}clone + Z_{trial}trial + Z_{set(loc:trial)}set + Z_{rep(set)}rep + Z_{nirsrep(rep)}nirsrep + \varepsilon. \quad \text{.....} \quad \text{eq.1.}$$

Where y_{ijk} = raw phenotypic observations; μ = population mean; loc = fixed effect for location; $Z_{clone}clone$ = random effects for clone: $clone \sim N(0, I\sigma^2_{clone})$; $Z_{trial}trial$ = random effect for trial: trial ~ N(0, $I\sigma^2_{trial}$); $Z_{set(loc:trial)}set$ = random effect for set nested in trial and location: set ~ N(0, $I\sigma^2_{set}$); $Z_{rep(set)}rep$ = random effect of clone replication nested in set: rep ~ N(0, $I\sigma^2_{rep}$); $Z_{nirsrep(rep)}nirsrep$ =

random effect of sample replications nested in clone replications: nirsrep ~ N(0, $I\sigma^2_{nirsrep}$*); and* ε *= error term:* $\varepsilon \sim N(0, I\sigma^2_{\varepsilon})$.

The resulting BLUPS from Eq. 1 were further used to fit a single trait GS model and for comparing genomic estimated breeding values to derive prediction accuracies. Genomic estimated breeding values (GEBVs) for the clones were extracted from a genomic BLUPS (GBLUPS) model using a linear and non-linear model procedures by fitting a mixed model:

$$y = X\beta + Zu + \varepsilon$$

u ~ $N(0, \sigma^2_u K)$ and $\varepsilon \sim N(0, I\sigma^2_{\varepsilon})$

where y = response vector for each trait and in this case, the BLUPS from Eq. 1; β = vector of fixed effects for the overall mean with the design matrix X; u is a vector of random additive genomic effects with the design matrix Z and K is the additive genomic relationship matrix generated from SNPs.

The one-stage alternative was fitted using the raw phenotypes from different locations, trials, sets and replications. Due to the unbalanced nature of the data, only the main fixed effects of location, trial and sets were used for ease of cross-validation. The design matrix for these environmental covariates (location, trial and set) and the design matrix for the markers were used together as the independent variables in the RF model. The linear model using GBLUP, for both one and two stage models, was carried out with *sommer* while the non-linear Random forest model was carried out with *randomForest* packages in R (Covarrubias-Pazaran 2016; Breiman 2001; Svetnik et al. 2003).

GS models: Multi-trait (MT) GS predictions – two and one-stage approaches:

Other than the multiple traits response in multi-traits model, the one- and two-stage approaches in multi-traits approach were similar to the one- and two-stage approaches of single trait models in terms of fixed and random variables. The two-stage MT genomic estimated breeding values for the clones was defined as:

$$Y = X\beta + Zu + \varepsilon$$

Where Y = response matrix of the nine carotenoids and in this case, the BLUPS derived from Eq. 1, X = design matrix for fixed effects, β is the matrix of fixed effects coefficients (location, training, set and replication effects), Z and ε are independent variable with N (0, V_Z, K) and N (0, Vε, I_n) with K as the additive genomic relationship matrix for the clones generated from SNPs. The MT genomic estimated breeding values were derived using the EMMREML package in R (Akdemir and Okeke 2015; R Core Team 2017).

Both ST and MT prediction accuracies were obtained from using a k-fold cross-validation scheme (Kohavi, 1995). The entire dataset was divided into training and testing set on a ratio of 80:20. The training set was used to estimate marker effects for predictions while the estimated marker effects were used to predict the breeding values of the testing set. The prediction accuracies were derived as a correlation between EGV and the predicted values using marker information (GEBV) in two-stage approaches and as the correlation between GEBV and phenotypes (adjusted for fixed effects) divided by the square root of trait heritability in one-stage methods (Hayes et al., 2015; .Wolc et al., 2011). Cross-validation of accuracies was iterated 30 times, which means that the reported accuracies were the mean of 30 iterations.

Traits correlations: Both phenotypic (r^2_P) and genotypic (r^2_G) correlations between the reported carotenoids were obtained using Pearson correlation. The r^2_P was performed on the raw phenotypes from different locations, trial, set and replications while the r^2_G was obtained by correlating the EGV from Eq. 1.

Genome-wide Association Analysis:

A genome-wide association analysis to identify genome-wide set of genetic variants in different individuals associated with the observed variants in the carotenoids was carried out using a mixed linear model:

$Y = X\beta + Zu + \varepsilon$ with var (y) = V = $A\sigma^2_g + I\sigma^2_\varepsilon$.

where y is an n × 1 vector of phenotypes with n being the sample size, β is a vector of fixed effects, g is an n × 1 vector of the total genetic effects of the individuals with g∼ N (0, $A\sigma^2_g$), A is interpreted as the genetic relationship matrix between individuals, I is an n × n identity matrix, and ε is a vector of residual effects with ε ∼ N (0, $I\sigma^2_\varepsilon$).

The GWAS for the carotenoids was carried out using GCTA tool (Yang et al. 2011). Markers were further filtered and 87380 SNPs with MAF > 5% were retained for the analysis.

Results and Discussion:

Vis/NIRS calibration for carotenoids:

Comparing calibration performances from the three calibration models on TCC - PLSR, PCR and RF, the correlation between the actual and predicted values within the calibration set was generally high (≥ 0.96) (Table 3.1). The same correlation value (0.99) was obtained from PLSR and RF while a correlation of 0.96 was obtained from

PCR. The correlation between the predicted values using models developed in the training set with the actual laboratory values in the testing set - r_{cv}, from the three methods were also very high – 0.97 in PLSR and 0.96 each for PCR and RF. The R^2_c from the methods was highest in PLSR (0.93), lowest in RF (0.87) and 0.92 with PCR. Similarly, the root mean error obtained in the training set vs testing set was highest in RF (2.77) and lowest in PLS (2.23) while PCR was 2.32.

Table 3.1: Calibration performance of the three calibration models - PLSR, PCR and RF for TCC.

Model/Stat.	PLS	PCR	RF
r_c	0.99	0.96	0.99
r_{cv}	0.97	0.96	0.96
R^2_c	0.93	0.92	0.87
RMSE	2.23	2.32	2.77

From the calibration result, the three models especially based on RMSE and R^2_c, were robust for quantitative analyses for TCC (Fox et al. 2012), although PLSR model had an average better calibration performance than the other two models. The result agrees with other studies where PLSR has been identified as the choice model for many calibration processes. It could be attributed to its ability in reducing the complexity of models using fewer principal components that contain more related information (N. Cao 2013; Shenk and Westerhaus 1991; Wold, Sjöström, and Eriksson 2001). Similarly, the use RF has been reported in recent multivariate calibration studies, and it has been argued to have a better or in some cases, comparable to PLSR especially with noisy

datasets (Ghasemi and Tavakoli 2013; Lee et al. 2012). The number of trees has been reported to affect calibration performance where trees less than 500 increase error fluctuations while higher values up to 4800 can potentially stabilize error fluctuations in RF models (Lee et al. 2012). The advantage of using nonlinear models like RF is to account for nonlinear relationships between variables (Lee et al. 2012; Ghasemi and Tavakoli 2013). Most importantly, the advantage of RF over PLS and PCR from this study was in overcoming unrestricted predictions from linear models leading to the prediction of negative values, especially while working on traits with extreme low values. While PLS and PCR had unconstrained prediction limits being linear models (Boerner 2017), the use of RF had constraints that ensured non-negative predictions since the predictions from RF are done through averaging the results obtained in several trees (Qi 2012; Breiman, L., Friedman, J.H., Olshen, R.A., and Stone 1984). One way of handling negative predictions where the expected outcome should be positive in developing calibration is to eliminate the negative values. Another option is to transform the predicted values, a measure that can further complicate the interpretation of the data developed from the factor components of linear models which lack actual physical meanings (Ghasemi and Tavakoli 2013; Vinet and Zhedanov 2011; Wold, Sjöström, and Eriksson 2001). On the other hand, nonlinear models maintain the original unit of the data (Ghasemi and Tavakoli 2013). Consequently, we used RF for the calibration and analyses of other carotenoids in this study.

Analysis of NRCRI TP for carotenoids using RF:

Over 4000 spectra from NRCRI obtained from the two trials (TP1 and TP2), three locations – Umudike, Otobi and Kano, different sets and replications were analyzed

using the RF calibration model and the calibration procedure as described earlier. Reported calibration performance for the carotenoids using RF on training and testing sets include - r_c, r_{cv}, R^2_c and RMSE (Table 3.2). The final models used in analyzing the NRCRI spectral data for the carotenoids were fitted by combining the training and testing datasets in order to maximize the number of calibration samples.

Table 3.2: Calibration statistics of the portable Vis/NIRS spectra analyzed using RF for carotenoids using calibration set from CIAT in 2016.

Model	Stat.	TCC	AC	ATBC	LUT	VIO	9CBC	13CBC	15CBC	PHY
Cal.	$\mathbf{r_c}$	0.96	0.87	**0.97**	0.77	0.79	0.90	0.92	0.92	**0.66**
	$\mathbf{r_{cv}}$	0.96	0.86	**0.97**	0.73	0.77	0.89	0.91	0.91	**0.62**
	$\mathbf{R^2_c}$	0.89	0.64	**0.92**	0.66	0.59	0.76	0.75	0.80	**0.28**
	RMSE	2.65	0.01	1.6	0.32	0.14	0.26	0.38	0.06	2.9
	$\mathbf{n_t}$	132	59	132	84	132	132	132	131	71
Final	$\mathbf{R^2_c}$	0.90	0.72	0.93	**0.56**	**0.58**	0.78	0.80	0.82	**0.27**
	$\mathbf{RMSE_c}$	2.51	0.01	1.6	**0.33**	0.14	0.26	0.33	0.06	2.8
	$\mathbf{n_f}$	173	76	173	109	173	173	173	173	91

r_c = correlation between predicted and actual values in training set; r_{cv} = correlation between predicted and actual values in testing set; R^2_c = R^2 for calibration; RMSE = root mean square error; n_t = number of training set; n_f = number of final calibration set – combined training and testing sets.

The result of the calibration including the individual carotenoids showed that the correlation between actual and predicted values within the training set ranged from 0.66 in PHY to 0.97 in ATBC. Similarly, the correlation between the training and testing set ranged from 0.62 in PHY to 0.97 in ATBC (Table 3.2). The R^2_c was highest in ATBC (0.92) and low in PHY (0.28). Apart from PHY, the R^2_c for other traits were generally above 0.6. The RMSE was highest in PHY (2.9) and lowest in AC (0.01). The same

calibration parameters were reported for the final model with a combined set of training and testing sets (Table 3.2). There was on average, improvement in calibration performance for most traits with the increase in the number of samples (n_t to n_f) used for calibration (Table 3.2).

Generally, the R^2_c and the correlations between the true and predicted values between training and testing set of the calibration models for the carotenoids were very high and should be useful for screening and quantification purposes (Fox et al. 2012). Typically, R^2 of 0.50 has been described as being useful for the discrimination of concentrations, R^2 between 0.60–0.82 is useful for screening and quantification, between 0.83–0.90 is important in most applications, while between 0.92–0.96 is useful in most applications especially in quality assurance, and values above 0.98 is important for all applications (Fox et al. 2012; Cai et al. 2012). The calibration performance in PHY was generally lower than other traits. Comparing R^2 values between the reduced model with training set only and the final model developed with a combined set of training and testing sets, increase in the number of calibration samples could potentially improve calibration performance since there was increase in R^2 with reduction in prediction error in most of the traits (Table 3.2). Given that there are slight differences in maximum absorbance for the individual carotenoids (Tosato et al. 2016), it might be possible to use different pre-treatment and calibration models for each individual carotenoids.

Overall, as demonstrated in this study, RF is a useful tool in building a robust calibration model for carotenoids using spectra from the portable Vis/NIRS. It has been reported to be useful in overcoming over-fitting and valuable in handling nonlinear interferences,

noisy and outlier datasets in model development (Lee et al. 2012; Ghasemi and Tavakoli 2013; Svetnik et al. 2003).

Statistical summary and heritability of carotenoids from NRCRI:

From the analyzed NRCRI TP data, there was phenotypic variation in fresh cassava roots for the different carotenoids (Table 3.3). TCC values ranged from approximately 2 $\mu g g^{-1}$ to 15.39 $\mu g g^{-1}$ and an average of 4.72 $\mu g g^{-1}$ (fresh weight basis) was recorded in the study. Early attempt in using quantitative values to classify clones based on their TCC values grouped clones with TCC values up to 1.5–2.0 $\mu g g^{-1}$ into white parenchyma roots, TCC values ranging from 1.5 to 3.0 $\mu g g^{-1}$ - cream pulp roots, while TCC values above 3.0–3.5 $\mu g g^{-1}$ were classified as yellow roots. The current population therefore, had a considerable number of both white and yellow root clones similar to the population earlier used in GWAS studies for TCC (I. Y. Rabbi et al. 2017).

Table 3.3: Summary statistics and heritability of carotenoids from cassava.

Stat.	**TCC**	**AC**	**ATBC**	**LUT**	**VIO**	**9CBC**	**13CBC**	**15CBC**	**PHY**
Min.	2.20	0.05	0.53	0.14	0.22	0.23	0.28	0.05	3.68
Max.	15.39	0.07	10.18	1.45	0.61	1.15	1.44	0.26	8.99
Mean	4.72	0.06	1.58	0.25	0.33	0.44	0.56	0.10	5.41
SD	2.085	0.004	1.536	0.098	0.055	0.163	0.212	0.039	0.701
H^2	0.45	0.40	**0.46**	**0.38**	0.42	0.44	0.43	0.45	0.41
h^2	0.42	**0.30**	**0.44**	0.37	0.39	0.40	0.39	0.41	0.34

Phenotypic variability was equally recorded on the individual carotenoid components ranging from 0.53 μgg^{-1} to 10.18 μgg^{-1} with a mean of 1.58 μgg^{-1} for example, in ATBC. AC had a narrow range of variability - 0.05 μgg^{-1} to 0.07 μgg^{-1} with a mean of 0.06 μgg^{-1} and standard deviation of 0.004 among other component carotenoids (Table 3.3).

The broad and narrow sense heritability were moderate and ranged from 0.38 in LUT to 0.46 in ATBC (broad-sense) and 0.30 in AC to 0.44 in ATBC (narrow-sense). The broad and narrow sense heritability for TCC was recorded at 0.45 and 0.42, respectively. In general, many studies have demonstrated that TCC is a highly heritable trait (Esuma et al. 2016; H. Ceballos et al. 2013; I. Y. Rabbi et al. 2017; Morillo C et al. 2013) which permits the use of rapid cycling recurrent selection approach in increasing the trait (Iglesias et al. 1997; H. Ceballos et al. 2013). Previous studies on the inheritance of carotenoids have focused mainly on TCC, but the use of the portable Vis/NIRS in quantifying both TCC and the corresponding components is a great milestone in phenotyping large population of clones and understanding the inheritance as well as designing the best breeding strategy in improving these traits. This study therefore, is very relevant in the current effort in addressing VAD that is prevalent in many regions of the world (Chávez et al. 2005; Pillay et al. 2014) through bio-fortification of major staple foods, utilizing the natural genetic diversity observed in breeding programs for these traits (H. Ceballos et al. 2013; Bouis et al. 2011).

Carotenoids correlations:

Phenotypic correlations observed among the carotenoid components from this study were generally high and positive (Table 3.4). The phenotypic correlations between TCC and the individual components showed very high and positive correlations between

TCC and 15CBC (0.98), ATBC (0.97), 9CBC and 13CBC (0.91), VIO (0.79) and AC (0.75). However, LUT and PHY had a low but positive correlation (0.22) with TCC (Table 3.4). Among the carotenoid components, very high and positive phenotypic correlation was observed between 9CBC and 13CBC (appr. 1), and 0.97 between 15CBC and 9CBC as well as 13CBC. Low but positive correlations were observed among some carotenoids, the lowest observed between PHY and 13CBC (0.01). Other low but positive associations were observed between PHY and 15CBC (0.14), LUT and AC (0.15), ATBC (0.19) among others. Low and negative associations were observed between PHY and LUT (-0.14) as well as between PHY and 9CBC (-0.01). Most of the PVAC including ATBC, AC, 9CBC, 13CBC and 15CBC were highly and positively correlated (Table 3.4).

Table 3.4: Phenotypic correlation of cassava carotenoids

	TCC	AC	ATBC	LUT	VIO	9CBC	13CBC	15CBC	PHY
TCC	1	0.75	**0.97**	0.22	0.79	0.91	0.91	**0.98**	0.22
AC		1	0.68	0.15	0.76	0.66	0.68	0.74	0.66
ATBC			1	0.19	0.68	0.83	0.82	0.93	0.22
LUT				1	0.51	0.25	0.26	0.23	**-0.14**
VIO					1	0.87	0.87	0.84	0.18
9CBC						1	**1**	**0.97**	**-0.01**
13CBC							1	**0.97**	0.01
15CBC								1	0.14
PHY									1

Similarly, very high and positive genotypic correlations were recorded between TCC and most of the **β**-carotene isomers – ATBC, 9CBC, 13CBC and 15CBC (>0.91) as well as AC (0.76) and VIO (0.82) (Table 3.5). Low but positive correlations were observed between TCC and LUT (0.29) as well as PHY (0.2). The genotypic correlations among the other carotenoids components were similar to the reported phenotypic correlations. However, additional low and negative genotypic correlation was observed between PHY and 13CBC (-0.01) in addition to a similar low and negative associations between PHY and LUT (-0.13) as well as 9CBC (-0.04) recorded in the phenotypic correlations (Table 3.5).

Table 3.5: Genotypic correlation of carotenoids

	TCC	AC	ATBC	LUT	VIO	9CBC	13CBC	15CBC	PHY
TCC	1	0.76	0.96	0.29	0.82	0.91	0.91	0.98	0.2
AC		1	0.65	0.18	0.78	0.66	0.69	0.74	0.68
ATBC			1	0.24	0.68	0.83	0.82	0.93	0.15
LUT				1	0.56	0.28	0.3	0.28	-0.13
VIO					1	0.87	0.88	0.85	0.17
9CBC						1	1	0.97	-0.04
13CBC							1	0.97	-0.01
15CBC								1	0.1
PHY									1

The phenotypic and genotypic correlations reported in this study were possible due to the development of an advanced phenotyping protocol that allowed for the much needed partitioning of TCC into its various components (Ceballos et al., 2017). Understanding the relationship (phenotypic and genotypic) especially between TCC and its components is vital in assessing the amount of progress made so far or needed to be made in increasing TCC as well as its components, most especially, the vitamin A precursors in cassava roots. The high and positive phenotypic and genotypic relationships observed especially between TCC and ATBC as well as other vitamin A precursory carotenoids is quite encouraging and suggest that these traits could be improved concurrently. Another association of interest was the positive phenotypic and genotypic association between LUT and ATBC even though the values were low (0.19 and 0.24, respectively) (Tables 3.4 and 3.5). So much emphasis has not been placed on LUT and probably because it is not a vitamin A precursor, however, it is a very important component of the macular pigment in the eyes and its deficiency is closely associated with some eye-related problems (Kim et al. 2010; Bechoff et al. 2015; Krinsky and Johnson 2005). An important consideration is the low and negative associations between PHY and some of the **β**-carotene components as well as LUT. PHY is a colorless carotenoid, and its synthesis is the first carotenoid precursor in the biosynthetic pathway of other carotenoids (Meléndez-Martínez et al. 2015; H. Cao et al. 2012). Its synthase has been reported as the limiting stage for carotenoid biosynthesis and regulation (Paine et al. 2005; Welsch et al. 2010a; Maass et al. 2009). The result of the correlations, especially genotypic, are important in designing strategies for the improvement of these traits. Similar high and positive correlations reported in this study,

particularly, between AC and PHY as well as between AC and ATBC have been reported in carrot roots (Fernandes Santos, Senalik, and Simon 2005).

Genome-wide association studies (GWAS):

GWAS is significant in the identification of genomic regions that are associated with variations in different carotenoid components. From this study, there was a total of 42 unique significant markers (at P values less than the 5% Bonferroni threshold) associated with variation in TCC and other carotenoids. Most of the significant markers were associated with variation in more than one trait (Table 3.6). There were no significant hit for AC or PHY from this study (Figure 3.1). The major regions associated with variations in different carotenoid components are within chromosomes 1, 2, 4, 13, 14 and 15.

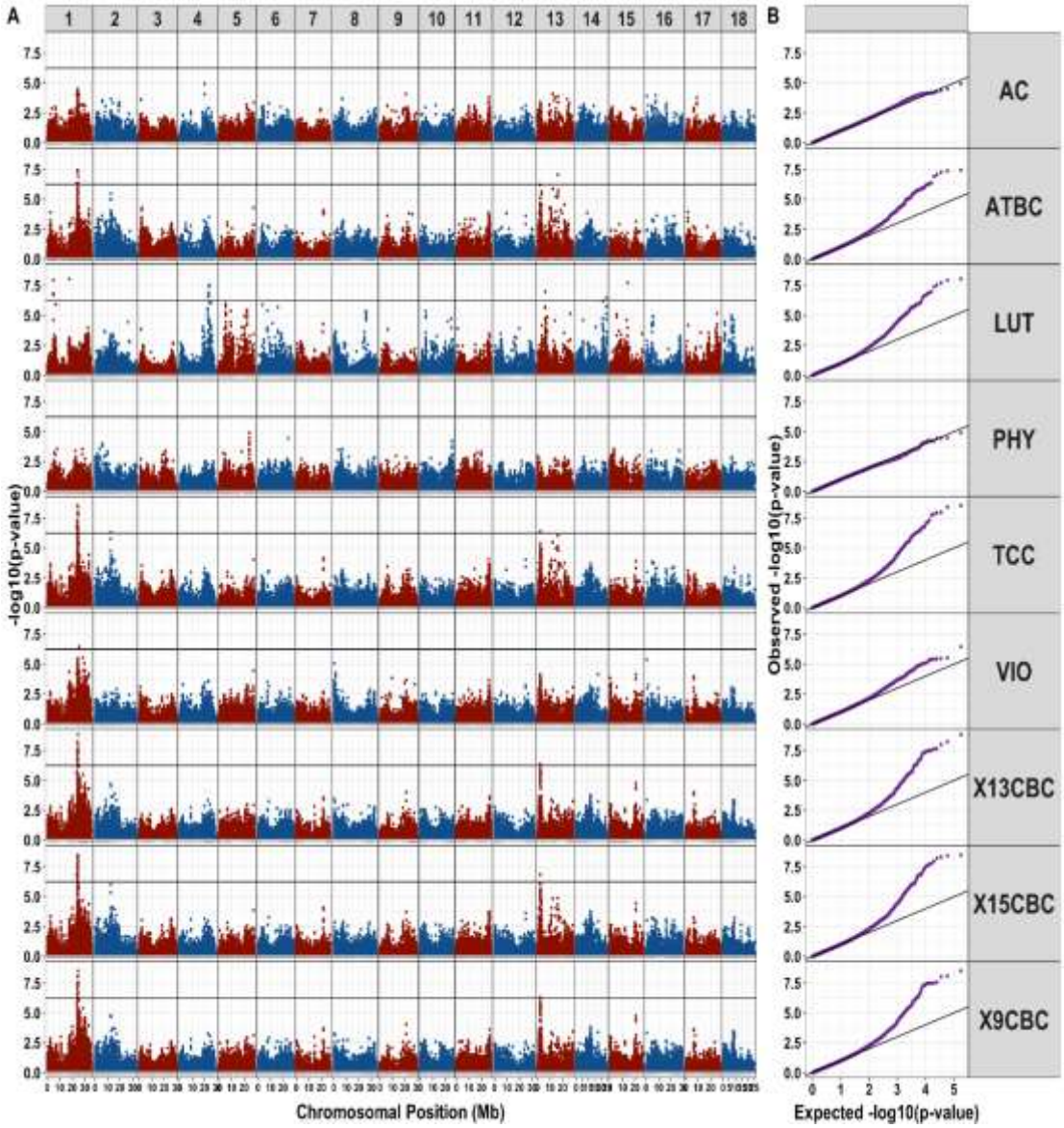

Figure 3.1: The Manhattan (A) and QQ (B) plots from the GWAS of the different carotenoids components. X13CBC = 13-cis beta-carotene; X15CBC = 15-cis beta-carotene and X9CBC = 9-cis beta-carotene.

A total of 20 markers were significant for variation in TCC, and 17 of those markers tagged a major peak located between 23.386Mbp to 24.709Mbp on chromosome 1. A single marker tagged another peak around 12.739Mbp on Chromosome 2 (P-value =

4.71 x 10^{-7}) and the remaining two markers approximately around 21.85Mbp (P-value = 4.34 x 10^{-7}) tagged the other peak on chromosome 13 (Table 3.6). Interestingly, similar regions tagged by almost the same markers for variation in TCC were equally significant for variations in ATBC, 9CBC, 13CBC and 15CBC. In addition, there was a nearby peak at 25.427Mbp tagged by one marker (P-value = 3.50 x 10^{-7}) significant for both variation in 13CBC and VIO. On the other hand, five regions were associated with variation in LUT on chromosome 1 tagged by four significant markers of which three were approximately localized between 4.81Mbp and 4.86Mbp and the remaining marker around 17.48Mbp. Five markers tagged a peak around 22.54Mbp to 23.69Mbp on chromosome 4 and a marker each on chromosome 13 (6.09Mbp and P-values = 1.04 x 10^{-7}), chromosome 14 (24.24Mbp and P-values = 3.28 x 10^{-7}), and chromosome 15 (14.17Mbp and P-values = 1.86 x 10^{-8}).

Table 3.6: Markers with genome-wide association significance for carotenoids in cassava roots.

Trait	Marker	Chr	Pos. (Mb)	Allele	Freq.	SNP Effect	p-value
ATBC	1_23386060	1	23.39	T/C	0.28	0.385	4.94E-07
ATBC	1_24105199	1	24.11	C/A	0.32	0.438	3.69E-08
ATBC	1_24181156	1	24.18	A/G	0.47	0.328	5.51E-08
ATBC	1_24611696	1	24.61	T/G	0.31	0.432	4.36E-08
ATBC	1_24636113	1	24.64	G/A	0.31	0.400	4.36E-07
ATBC	1_24709749	1	24.71	T/A	0.31	0.416	1.23E-07
ATBC	13_16164208	13	16.16	T/A	0.05	0.592	8.98E-08
LUT	1_4814833	1	4.81	G/T	0.05	0.031	1.98E-07
LUT	1_4815271	1	4.82	T/A	0.06	0.027	1.42E-07
LUT	1_4857197	1	4.86	C/A	0.06	0.032	1.13E-08
LUT	1_17482176	1	17.48	A/G	0.06	0.032	8.78E-09
LUT	4_22536069	4	22.54	T/G	0.06	0.030	2.62E-07
LUT	4_22974535	4	22.97	G/A	0.06	0.030	1.25E-07
LUT	4_23362863	4	23.36	A/G	0.05	0.034	3.96E-08
LUT	4_23367235	4	23.37	C/T	0.06	0.030	2.01E-07
LUT	4_23693408	4	23.69	G/A	0.05	0.034	2.78E-08
LUT	13_6088827	13	6.09	T/A	0.09	0.025	1.04E-07
LUT	14_24239911	14	24.24	T/C	0.05	0.031	3.28E-07
LUT	15_14171698	15	14.17	G/A	0.05	0.031	1.86E-08
TCC	1_23386060	1	23.39	T/C	0.28	0.512	1.77E-07

TCC	1_24105199	1	24.11	C/A	0.32	0.606	2.88E-09
TCC	1_24117585	1	24.12	C/G	0.35	0.488	4.77E-08
TCC	1_24121247	1	24.12	T/C	0.36	0.464	8.54E-08
TCC	1_24139256	1	24.14	C/T	0.48	0.449	2.71E-07
TCC	1_24140688	1	24.14	G/C	0.48	0.473	7.08E-08
TCC	1_24159583	1	24.16	T/C	0.36	0.443	5.05E-07
TCC	1_24181156	1	24.18	A/G	0.47	0.442	1.04E-08
TCC	1_24238287	1	24.24	T/C	0.45	0.450	3.92E-07
TCC	1_24239005	1	24.24	G/A	0.44	0.476	1.58E-07
TCC	1_24272965	1	24.27	G/C	0.41	0.465	5.51E-07
TCC	1_24315496	1	24.32	G/A	0.40	0.443	1.97E-07
TCC	1_24611696	1	24.61	T/G	0.31	0.597	3.80E-09
TCC	1_24614646	1	24.61	C/A	0.35	0.435	2.57E-07
TCC	1_24636113	1	24.64	G/A	0.31	0.572	1.80E-08
TCC	1_24653227	1	24.65	C/G	0.32	0.570	1.14E-08
TCC	1_24709749	1	24.71	T/A	0.31	0.572	1.46E-08
TCC	2_12738969	2	12.74	A/G	0.06	0.749	4.71E-07
TCC	13_2185403	13	2.19	G/C	0.15	0.591	4.34E-07
TCC	13_2185406	13	2.19	A/T	0.15	0.591	4.34E-07
VIO	1_25426915	1	25.43	T/G	0.36	0.010	3.50E-07
13CBC	1_24105199	1	24.11	C/A	0.32	0.053	3.37E-08
13CBC	1_24117585	1	24.12	C/G	0.35	0.047	2.41E-08

Table 3.6 (Continued)

13CBC	1_24121247	1	24.12	T/C	0.36	0.048	5.81E-09
13CBC	1_24121295	1	24.12	T/C	0.36	0.043	1.92E-07
13CBC	1_24121316	1	24.12	T/A	0.36	0.043	1.81E-07
13CBC	1_24139256	1	24.14	C/T	0.48	0.045	6.23E-08
13CBC	1_24140688	1	24.14	G/C	0.48	0.044	1.01E-07
13CBC	1_24159583	1	24.16	T/C	0.36	0.046	4.51E-08
13CBC	1_24160008	1	24.16	G/A	0.35	0.043	2.88E-07
13CBC	1_24181156	1	24.18	A/G	0.47	0.040	3.85E-08
13CBC	1_24239005	1	24.24	G/A	0.44	0.043	5.52E-07
13CBC	1_24611696	1	24.61	T/G	0.31	0.055	1.01E-08
13CBC	1_24614646	1	24.61	C/A	0.35	0.045	2.30E-08
13CBC	1_24632970	1	24.63	T/C	0.37	0.039	5.29E-07
13CBC	1_24636113	1	24.64	G/A	0.31	0.053	3.40E-08
13CBC	1_24653227	1	24.65	C/G	0.32	0.057	1.44E-09
13CBC	1_24663824	1	24.66	G/C	0.35	0.042	1.78E-07
13CBC	1_24664143	1	24.66	A/G	0.35	0.045	4.89E-08
13CBC	1_24709749	1	24.71	T/A	0.31	0.053	2.97E-08
13CBC	1_25426915	1	25.43	T/G	0.36	0.040	5.71E-07
13CBC	13_2185403	13	2.19	G/C	0.15	0.056	4.47E-07
13CBC	13_2185406	13	2.19	A/T	0.15	0.056	4.47E-07

Table 3.6 (Continued)

13CBC	13_2208957	13	2.21	A/G	0.14	0.058	5.23E-07
15CBC	1_23386060	1	23.39	T/C	0.28	0.010	1.44E-07
15CBC	1_24105199	1	24.11	C/A	0.32	0.011	3.37E-09
15CBC	1_24117585	1	24.12	C/G	0.35	0.009	2.00E-08
15CBC	1_24121247	1	24.12	T/C	0.36	0.009	1.64E-08
15CBC	1_24121295	1	24.12	T/C	0.36	0.008	2.56E-07
15CBC	1_24121316	1	24.12	T/A	0.36	0.008	2.77E-07
15CBC	1_24139256	1	24.14	C/T	0.48	0.009	4.31E-08
15CBC	1_24140688	1	24.14	G/C	0.48	0.009	2.84E-08
15CBC	1_24159583	1	24.16	T/C	0.36	0.009	8.40E-08
15CBC	1_24160008	1	24.16	G/A	0.35	0.008	4.47E-07
15CBC	1_24181156	1	24.18	A/G	0.47	0.008	5.99E-09
15CBC	1_24224353	1	24.22	T/G	0.48	0.008	5.10E-07
15CBC	1_24239005	1	24.24	G/A	0.44	0.009	1.70E-07
15CBC	1_24611696	1	24.61	T/G	0.31	0.011	4.84E-09
15CBC	1_24614646	1	24.61	C/A	0.35	0.008	5.64E-08
15CBC	1_24636113	1	24.64	G/A	0.31	0.010	2.44E-08
15CBC	1_24653227	1	24.65	C/G	0.32	0.011	3.76E-09
15CBC	1_24664143	1	24.66	A/G	0.35	0.008	1.65E-07
15CBC	1_24709749	1	24.71	T/A	0.31	0.011	1.09E-08
15CBC	13_2185403	13	2.19	G/C	0.15	0.011	1.45E-07

Table 3.6 (Continued)

15CBC	13_2185406	13	2.19	A/T	0.15	0.011	1.45E-07
9CBC	1_24105199	1	24.11	C/A	0.32	0.043	3.50E-08
9CBC	1_24117585	1	24.12	C/G	0.35	0.038	3.32E-08
9CBC	1_24121247	1	24.12	T/C	0.36	0.038	9.33E-09
9CBC	1_24121295	1	24.12	T/C	0.36	0.034	3.38E-07
9CBC	1_24121316	1	24.12	T/A	0.36	0.034	2.88E-07
9CBC	1_24139256	1	24.14	C/T	0.48	0.036	3.83E-08
9CBC	1_24140688	1	24.14	G/C	0.48	0.036	6.47E-08
9CBC	1_24159583	1	24.16	T/C	0.36	0.036	1.05E-07
9CBC	1_24181156	1	24.18	A/G	0.47	0.032	3.47E-08
9CBC	1_24611696	1	24.61	T/G	0.31	0.044	8.18E-09
9CBC	1_24614646	1	24.61	C/A	0.35	0.036	2.76E-08
9CBC	1_24636113	1	24.64	G/A	0.31	0.042	4.96E-08
9CBC	1_24653227	1	24.65	C/G	0.32	0.045	3.01E-09
9CBC	1_24663824	1	24.66	G/C	0.35	0.034	1.93E-07
9CBC	1_24664143	1	24.66	A/G	0.35	0.036	4.86E-08
9CBC	1_24709749	1	24.71	T/A	0.31	0.042	3.14E-08
9CBC	13_2185403	13	2.19	G/C	0.15	0.045	5.04E-07
9CBC	13_2185406	13	2.19	A/T	0.15	0.045	5.04E-07

A reference was made to the cassava genome (v6.1) (Bredeson et al. 2016) in Phytozome (v12.1.6) (Goodstein et al. 2012) to identify annotated genes within a distance of ± 0.5 Mb of the genomic region occupied by the significant SNPs. The candidate gene *Manes.01G124200,* a phytoene synthase (PSY) gene known for increasing the accumulation of carotenoid in cassava roots (Esuma et al., 2016; Rabbi et al., 2017; Welsch et al., 2010b) and *Manes.01G001200* gene also associated with carotenoid biosynthesis (Goodstein et al. 2012; Bredeson et al. 2016), located within the genomic regions (~24.15 to 24.16 Mbp, forward and 25.21 to 25.48 Mbp, forward, respectively) were found around the regions of the significant markers on chromosome 1 which was associated with variation in TCC, ATBC, 9CBC, 13CBC, 15CBC and VIO. There were no known candidate genes found in the other regions associated with variation in the studied carotenoids on chromosomes 2, 4, 13, 14 and 15. However, many other genes with various biological functions were found around the regions of the significant markers. The findings here are in agreement with previous GWAS on TCC in cassava (Esuma et al. 2016; I. Y. Rabbi et al. 2017). The candidate gene, *Manes.01G124200* was reported as a single genomic region associated with variation in TCC, evaluated with quantitative TCC values obtained from HPLC using a panel of partial S1 and S2 generation inbreds generated from eight clones (Esuma et al. 2016). This gene was also identified using a collection of cassava clones representing diverse African germplasm phenotyped using an indirect color chart as well as Chromameter b* value for TCC (Rabbi et al., 2017).

Many studies have focused on understanding the inheritance pattern and the genomic regions associated with carotenoids using qualitative or quantitative measures of TCC.

Some of those studies identified a single major locus associated with carotenoid content variation using mapping populations and partial inbred lines (I. Rabbi et al. 2014; Welsch et al. 2010a; Esuma et al. 2016). However, the possibility of more than one associated locus has been suggested (Esuma et al. 2016; I. Y. Rabbi et al. 2017; Akinwale et al. 2010; Iglesias et al. 1997). More so, it is important to go beyond understanding the genes responsible for differences in white versus yellow clones to the underlying genetic control for quantitative variation in cassava roots (Hernán Ceballos et al. 2017). This study uncovered additional regions for variation in TCC and individual traits and identified markers that were significant for more than one carotenoid.

However, progress in quantitative analyses of clones in breeding programs for carotenoids could only be possible with improved phenotyping protocols. Lack of such protocols often limited most of the earlier studies to visual assessment of the differences in root parenchyma pigmentation intensity as an indirect measure for TCC. The use of HPLC which has the potential for quantitative assessment and dissection of TCC into its various components was adopted by programs but the output is generally low, restricting the number of samples that could be screened within a given period (Sánchez et al. 2014; Belalcazar, Dufour, Andersson, Pizarro, Luna, Londoño, Morante, Jaramillo, Pino, López-Lavalle, Davrieux, Talsma, and Ceballos 2016; Hernán Ceballos et al. 2017). It has been reported to lengthen harvesting time for a long period and thereby delaying the time to make breeding decisions and potentially introducing variations in the overall screening process with considerable impact on the assessment of some other quality traits such as DMC (Hernán Ceballos et al. 2017). Besides, the

use of HPLC demands certain technical knowledge and infrastructures to function effectively.

This study therefore, supports the use of alternative phenotyping methods offered by NIRS technology and represents one of the initial attempts in identifying genomic regions associated with individual carotenoids. Information derived from the study will help to expand breeding targets to account for the underlying genes controlling carotenoids accumulation in cassava roots.

Genomic predictions:

Although the idea was not strictly to compare prediction accuracies from different models, the result of the genomic predictions of carotenoids showed that on average, predictions using one-stage approach were higher in all the cases of single and multiple traits GBLUB and the use of RF than using two-stage method (Figure 3.2). The use of multi-traits GBLUP led to higher accuracies in most cases except in PHY than the use of single trait GBLUP model under the two-stage approach. Conversely, the use of the single trait model was higher than the multiple traits model under the one-stage approach. Overall, the prediction accuracies from non-linear model using RF was higher than the linear single trait or multiple traits predictions in both one-stage and two-stage models. On average prediction accuracies were high (approximately, 0.6) in TCC, ATBC, 9CBC, 13CBC and 15CBC; moderate in 0.47 in VIO and AC; about 0.35 in LUT and 0.2 in PHY.

There is an ongoing effort to use GS to reduce the breeding cycle of cassava and accelerate the rate of genetic gain for major traits (Wolfe et al. 2017; E. J. Oliveira et al. 2014; Okeke et al. 2017). The efficiency of GS in both plant and animal breeding has

already been demonstrated (Lorenz et al. 2011; X Zhang et al. 2015; Daetwyler et al. 2013) and its implementation in cassava holds many promises to the millions of people that depend on cassava for food and source of income. The use of the multiple traits model, which uses estimate of genetic and residual covariance in deriving GEBV for the traits of interest, has been demonstrated in GS (Jia and Jannink 2012; Okeke et al. 2017). On average, we obtained higher prediction accuracies using multi-trait GBLUP model in the two-stage approach for almost all the traits. In the one-stage approach, the advantage of multi-traits model was not realized which could be attributed to unaccounted environmental variations besides, heritability and trait correlation factors (Guo et al. 2014; Calus and Veerkamp 2011). The advantage of multi-traits over single trait models is fully realized where there is medium to high genetic correlations in the joint analyses of low and high heritable traits (Calus and Veerkamp 2011; Okeke et al. 2017).

Although the two-stage approach is generally used in GS due to its simplicity and computational efficiency, the one-stage approach is usually regarded as the gold standard since it effectively accounts for the entire variance–covariance structure of the observed data (Möhring and Piepho 2009; Smith, Cullis, and Thompson 2001; Schulz-Streeck, Ogutu, and Piepho 2013). With main factor effects, we observed slightly higher prediction accuracies using one-stage than two-stage approaches in all the cases of linear single trait, multi-traits and non-linear single traits models. Depending on the complexity of the dataset, model and number of markers and genotypes involved, the one-stage analysis can be computationally demanding. However, different forms of weighting have been suggested to minimize information loss in the two-stage approach

(Garrick, Taylor, and Fernando 2009; Schulz-Streeck, Ogutu, and Piepho 2013; Möhring and Piepho 2009).

Also, the use of non-linear GS model was on average higher than the linear single and multi-traits models across the one- and two-stage scenarios. Similar results have been widely reported (Crossa et al. 2014; Pérez-Rodríguez et al. 2012; Heslot et al. 2012) and it is relevant in capturing dominance and epistasis effects in predicting genome estimated total genetic value other than GEBV (Wolfe et al. 2017; Spindel et al. 2015). This is valuable for crops like cassava and rice where released varieties are clones and inbreds, respectively. The higher accuracies obtained from RF in the two-stage approach over the linear models could have resulted from non-additive genetic effects, whereas the higher accuracies from the one-stage method could be attributed to some non-genetic interactions captured by the model.

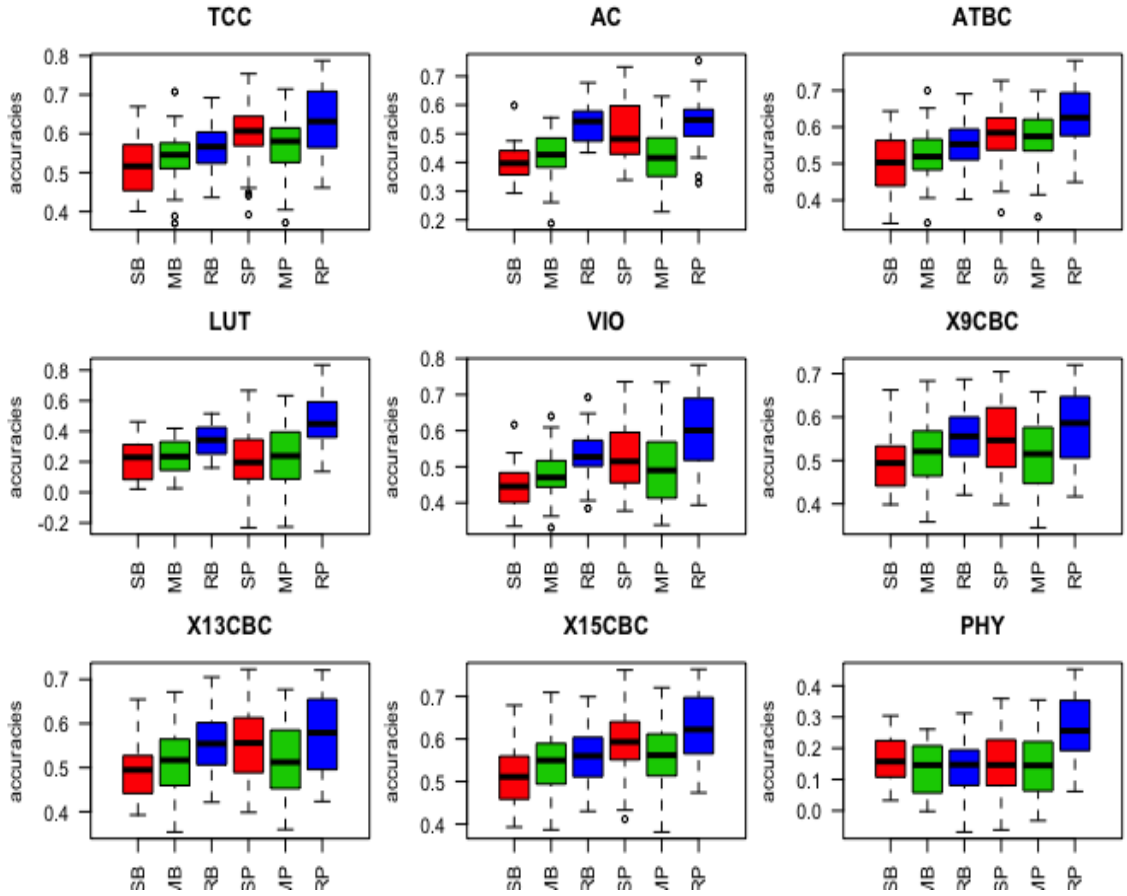

Figure 3.2: Genomic predictions for carotenoids. From left to right for each trait: SB = two-stage, single trait GBLUB; MB = two-stage multi-traits GBLUP; RB = two-stage, single trait RF; SP = one-stage, single trait GBLUP; MP = one-stage multi-traits GBLUP; RP = one-stage, single trait RF; X13CBC = 13-cis beta-carotene; X15CBC = 15-cis beta-carotene and X9CBC = 9-cis beta-carotene.

Conclusion

In order to make meaningful progress in the nutritional improvement of cassava, thorough assessment and partitioning of TCC into its various components is very valuable. Previously, HPLC has been useful in this regard but it is very slow and low throughput. However, NIRS provides an alternative for rapid and large-scale phenotyping in breeding and nutritional evaluations (Berardo et al. 2004; Teye, Huang, and Afoakwa 2013; Lebot et al. 2009; Sánchez et al. 2014). The introduction of portable NIRS versions offer additional flexibility for field-based analyses (Ikeogu et al. 2017). In developing an efficient calibration model, the choice of calibration models as well as pre-treatment methods are very important for extracting meaningful information relevant to the chemical constituent of interest from the spectral data, while correcting for external interferences mostly associated with light scattering, background noise and baseline drifts. From the result obtained in this study, RF could be a reliable tool in developing robust models especially where the relationship between dependent and independent variables are not necessarily linear. The improved phenotyping protocol enabling the quantitative and simultaneous evaluation of both TCC and the individual components provides a good foundation for re-defining breeding strategies in the improvement of these traits in cassava. NIRS clearly overcomes the challenges of conventional carotenoids quantification; it is fast and can be used for the simultaneous assessment of many other traits. Another important milestone in the use of NIRS for carotenoids quantification is the ability to use the device in tracking carotenoids concentration in cassava roots before and after processing. This is important to translate fresh weight into dry weight concentrations in the final products since the relationship

between carotenoids concentrations on fresh and dry weight basis is not always linear, and retention of carotenoids is highly dependent on processing method and clones among other things (Hernán Ceballos et al. 2017; Iglesias et al. 1997).

The positive and high phenotypic and genotypic associations observed in this study underscores the fact that any effort in increasing TCC could lead to increase in the individual components. However, there were low and negative associations among some of the individual carotenoids, which requires special consideration in designing an efficient breeding scheme. We verified and identified both common and unique regions associated with variations in TCC and most of the carotenoid components. The identified genes could be deployed in breeding programs and support the current effort in increasing the nutritional quality of cassava.

GS has been useful in predicting the genetic potential of individuals in many breeding programs. We have demonstrated that it could be useful in fast-tracking the quantitative improvement of carotenoids with good prediction accuracies. The heritability values, especially the narrow sense heritability, obtained from this study are sufficient enough in deriving GEBVs with useful accuracy in using GS to improve these traits as has been demonstrated in other traits and other species (Ly et al. 2013; B J Hayes et al. 2017; Ben J. Hayes et al. 2010; Wolfe et al. 2017). Similar to the development of calibration models, the use of non-linear GS models have potentials to capture non-linear underlying relationships between dependent and independent variables. It has implication in cassava breeding where such non-linear dominance and epistasis effects are beneficial in predicting total genetic values other than GEBV (Wolfe et al. 2017; Heslot et al. 2012). This is one of the initial attempts in dissecting the genetic

architecture of individual carotenoids in cassava breeding. The loci associated with carotenoids variation could be useful in designing MAS for these traits. Also, information from the GWAS analysis could be incorporated into GS to improve predictions for carotenoids content in the genetic background of other relevant agronomic traits (Wolfe et al. 2016; Spindel et al. 2015).

Although genotyping cost is drastically decreasing, it is still not so cheap for resource-limited breeding programs and to genotype large collection of individuals typical of early breeding generations. To reduce genotyping and classical phenotyping costs, there has been a growing interest in incorporating descriptors from NIRS in improving genomic predictions (B J Hayes et al. 2017). NIRS wavelength regions significant for some important traits could be targeted for candidate genes to be used in trait improvements. The implication of rapid phenotyping using NIRS in the implementation of GS in cassava as well as other crops include the reduction of phenotyping cost and time enabling the use of more individuals in the training population necessary for increasing genetic diversity and selection intensity and shortening the time in making breeding decisions.

REFERENCES

Akdemir, D., & Okeke, U. G. (2015). EMMREML: Fitting Mixed Models with Known Covariance Structures. *Https://Cran.r-Project.Org/Package=EMMREML.*, R package version 3.1. Retrieved from https://cran.r-project.org/web/packages/EMMREML/EMMREML.pdf

Akinwale, M. G., Aladesanwa, R. D., Akinyele, B. O., Dixon, A. G. O., & Odiyi, A. C. (2010). Inheritance of B-carotene in cassava (Manihot esculenta crantz). *International Journal of Genetics and Molecular Biology*, *2*(10), 198–201. Retrieved from http://www.academicjournals.org/ijgmb

Andersson, M. (2009). A comparison of nine PLS1 algorithms. *Journal of Chemometrics*, *23*(10), 518–529. http://doi.org/10.1002/cem.1248

Armstrong, G. A., & Hearst, J. E. (1996). Carotenoids 2: Genetics and molecular biology of carotenoid pigment biosynthesis. *The FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology*, *10*(2), 228–237. Retrieved from http://www.fasebj.org.proxy.library.cornell.edu/content/10/2/228.full.pdf

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. http://doi.org/10.18637/jss.v067.i01

Bechoff, A., Chijioke, U., Tomlins, K. I., Govinden, P., Ilona, P., Westby, A., & Boy, E. (2015). Carotenoid stability during storage of yellow gari made from biofortified cassava or with palm oil. *Journal of Food Composition and Analysis*, *44*, 36–44. http://doi.org/10.1016/j.jfca.2015.06.002

Belalcazar, J., Dufour, D., Andersson, M. S., Pizarro, M., Luna, J., Londoño, L., … Ceballos, H. (2016). High-throughput phenotyping and improvements in breeding cassava for increased carotenoids in the roots. *Crop Science*, *56*(6), 2916–2925. http://doi.org/10.2135/cropsci2015.11.0701

Belalcazar, J., Dufour, D., Andersson, M. S., Pizarro, M. M., Luna, J., Londoño, L., … Ceballos, H. H. (2016). High-throughput phenotyping and improvements in breeding cassava for increased carotenoids in the roots. *Crop Science*, *56*(6), 2916–2925. http://doi.org/10.2135/cropsci2015.11.0701

Berardo, N., Brenna, O. V, Amato, A., Valoti, P., Pisacane, V., & Motto, M. (2004). Carotenoids concentration among maize genotypes measured by near infrared reflectance spectroscopy (NIRS). *Innovative Food Science and Emerging Technologies*, *5*(3), 393–398. http://doi.org/10.1016/j.ifset.2004.03.001

Biau, G. (2010). Analysis of a Random Forests Model. Retrieved from https://bigdata.unl.edu/documents/ASA_Workshop_Materials/Analysis of a Random Forests Model.pdf

Boerner, V. (2017). On marker-based parentage verification via non-linear optimization. *Genetics Selection Evolution*, *49*(1), 50. http://doi.org/10.1186/s12711-017-0324-3

Bouis, H. E., Hotz, C., McClafferty, B., Meenakshi, J. V., & Pfeiffer, W. H. (2011). Biofortification: A new tool to reduce micronutrient malnutrition. *Food and Nutrition Bulletin*, *32*(1 SUPPL.). http://doi.org/10.1177/15648265110321S105

Bredeson, J. V, Lyons, J. B., Prochnik, S. E., Wu, G. A., Ha, C. M., Edsinger-

Gonzales, E., … Rokhsar, D. S. (2016). Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nature Biotechnology*, *34*(5), 562–570. http://doi.org/10.1038/nbt.3535

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C. J. (1984). *Classification and Regression Trees*. *Wadsworth & Brooks/Cole Advanced Books & Software* (Vol. 1).

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. http://doi.org/10.1023/A:1010933404324

Browning, B. L., & Browning, S. R. (2008). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American Journal of Human Genetics*, *84*(2), 210–223. http://doi.org/10.1016/j.ajhg.2009.01.005

Cai, R., Wang, S., Meng, Y., Meng, Q., & Zhao, W. (2012). Rapid quantification of flavonoids in propolis and previous study for classification of propolis from different origins by using near infrared spectroscopy. *Analytical Methods*, *4*(8), 2388–2395. http://doi.org/10.1039/c2ay25184a

Calus, M., & Veerkamp, R. (2011). Accuracy of multi-trait genomic selection using different methods. *Genet. Select. Evol.*, *43*(1), 26. http://doi.org/10.1186/1297-9686-43-26

Cao, H., Zhang, J., Xu, J., Xu, J., Ye, J., Yun, Z., … Deng, X. (2012). Comprehending crystalline β-carotene accumulation by comparing engineered cell models and the natural carotenoid-rich system of citrus. *Journal of Experimental Botany*, *63*(12), 4403–4417. http://doi.org/10.1093/jxb/ers115

Cao, N. (2013). Calibration optimization and efficiency in near infrared spectroscopy, 183. Retrieved from http://lib.dr.iastate.edu/etd/13199/

Castenmiller, J. J. M., & West, C. E. (1998). BIOAVAILABILITY AND BIOCONVERSION OF CAROTENOIDS. *Annual Review of Nutrition*, *18*(1), 19–38. http://doi.org/10.1146/annurev.nutr.18.1.19

Ceballos, H., Davrieux, F., Talsma, E. F., Belalcazar, J., Chavarriaga, P., & Andersson, M. S. (2017). Carotenoids in Cassava Roots. In *Carotenoids* (Vol. 3). http://doi.org/10.5772/intechopen.68279

Ceballos, H., Morante, N., Sánchez, T., Ortiz, D., Aragón, I., Chávez, A. L., … Dufour, D. (2013). Rapid cycling recurrent selection for increased carotenoids content in cassava roots. *Crop Science*, *53*(6), 2342–2351. http://doi.org/10.2135/cropsci2013.02.0123

Chávez, A. L., Sánchez, T., Jaramillo, G., Bedoya, J. M., Echeverry, J., Bolaños, E. A., … Iglesias, C. A. (2005). Variation of quality traits in cassava roots evaluated in landraces and improved clones. *Euphytica*, *143*(1–2), 125–133. http://doi.org/10.1007/s10681-005-3057-2

Chen, J., & Wang, X. Z. (2001). A New Approach to Near-Infrared Spectral Data Analysis Using Independent Component Analysis. *Journal of Chemical Information and Computer Sciences*, *41*(4), 992–1001. http://doi.org/10.1021/ci0004053

Covarrubias-Pazaran, G. (2016). Genome-Assisted prediction of quantitative traits using the r package sommer. *PLoS ONE*, *11*(6), e0156744.

http://doi.org/10.1371/journal.pone.0156744
Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to Support Vector Machines*. *History* (Vol. 47). http://doi.org/0521780195
Crossa, J., Pérez, P., Hickey, J., Burgueño, J., Ornella, L., Cerón-Rojas, J., … Mathews, K. (2014). Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity*, *112*(1), 48–60. http://doi.org/10.1038/hdy.2013.16
Daetwyler, H. D., Calus, M. P. L., Pong-Wong, R., de los Campos, G., & Hickey, J. M. (2013). Genomic prediction in animals and plants: Simulation of data, validation, reporting, and benchmarking. *Genetics*. http://doi.org/10.1534/genetics.112.147983
de Oliveira, E. J., de Resende, M. D. V., da Silva Santos, V., Ferreira, C. F., Oliveira, G. A. F., da Silva, M. S., … Aguilar-Vildoso, C. I. (2012). Genome-wide selection in cassava. *Euphytica*, *187*(2), 263–276. http://doi.org/10.1007/s10681-012-0722-0
Esuma, W., Herselman, L., Labuschagne, M. T., Ramu, P., Lu, F., Baguma, Y., … Kawuki, R. S. (2016). Genome-wide association mapping of provitamin A carotenoid content in cassava. *Euphytica*, *212*(1), 97–110. http://doi.org/10.1007/s10681-016-1772-5
FAO/WHO, F. and A. O. /World H. O. (1998). Vitamin and mineral requirements in human nutrition. Retrieved from http://www.fao.org/ag/humannutrition/36659-04427f866c8b2539d8e47d408cad5f3f9.pdf
Fernandes, S. B., Dias, K. O. G., Ferreira, D. F., & Brown, P. J. (2017). Efficiency of multi-trait, indirect, and trait-assisted genomic selection for improvement of biomass sorghum. *Theoretical and Applied Genetics*, 1–9. http://doi.org/10.1007/s00122-017-3033-y
Fernandes Santos, C. A., Senalik, D., & Simon, P. W. (2005). Path analysis suggests phytoene accumulation is the key step limiting the carotenoid pathway in white carrot roots. *Genetics and Molecular Biology*, *28*(2), 287–293. http://doi.org/10.1590/S1415-47572005000200019
Fox, G. P., O'Donnell, N. H., Stewart, P. N., & Gleadow, R. M. (2012). Estimating hydrogen cyanide in forage sorghum (Sorghum bicolor) by near-infrared spectroscopy. *Journal of Agricultural and Food Chemistry*, *60*(24), 6183–6187. http://doi.org/10.1021/jf205030b
Garrick, D. J., Taylor, J. F., & Fernando, R. L. (2009). Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genetics Selection Evolution*, *41*(1). http://doi.org/10.1186/1297-9686-41-55
Ghasemi, J. B., & Tavakoli, H. (2013). Application of random forest regression to spectral multivariate calibration. *Analytical Methods*, *5*(7), 1863. http://doi.org/10.1039/c3ay26338j
Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., … Rokhsar, D. S. (2012). Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research*, *40*(D1). http://doi.org/10.1093/nar/gkr944
Guo, G., Zhao, F., Wang, Y., Zhang, Y., Du, L., & Su, G. (2014). Comparison of single-trait and multiple-trait genomic prediction models. *BMC Genetics*, *15*. http://doi.org/10.1186/1471-2156-15-30

Habier, D., Fernando, R. L., & Dekkers, J. C. M. (2009). Genomic selection using low-density marker panels. *Genetics*, *182*(1), 343–353. http://doi.org/10.1534/genetics.108.100289

Hammond, B. R., & Renzi, L. M. (2013). Carotenoids. *Advances in Nutrition: An International Review Journal*, *4*(4), 474–476. http://doi.org/10.3945/an.113.004028

Hayes, B. J., Donoghue, K. A., Reich, C., Mason, B., Herd, R. M., & Arthur5, P. F. (2015). Genomic Estimated Breeding Values for Methane Production in. *Proc. Assoc. Advmt. Breed. Genet.*, 118–121. Retrieved from http://www.aaabg.org/aaabghome/AAABG21papers/Hayes21118.pdf

Hayes, B. J., Panozzo, J., Walker, C. K., Choy, A. L., Kant, S., Wong, D., … Spangenberg, G. C. (2017). Accelerating wheat breeding for end-use quality with multi-trait genomic predictions incorporating near infrared and nuclear magnetic resonance-derived phenotypes. *Theoretical and Applied Genetics*, *1*(0123456789). http://doi.org/10.1007/s00122-017-2972-7

Hayes, B. J., Pryce, J., Chamberlain, A. J., Bowman, P. J., & Goddard, M. E. (2010). Genetic Architecture of Complex Traits and Accuracy of Genomic Prediction: Coat Colour, Milk-Fat Percentage, and Type in Holstein Cattle as Contrasting Model Traits. *PLoS Genetics*, *6*(9), e1001139. http://doi.org/10.1371/journal.pgen.1001139

Hebuterne, X., Wang, X. D., Johnson, E. J., Krinsky, N. I., & Russell, R. M. (1995). Intestinal absorption and metabolism of 9-_cis_-  -carotene _in vivo: biosynthesis of 9-cis_-retinoic acid. *J.Lipid Res.*, *36*, 1264–1273. Retrieved from http://www.jlr.org/content/36/6/1264.full.pdf

Heslot, N., Yang, H.-P., Sorrells, M. E., & Jannink, J.-L. (2012). Genomic Selection in Plant Breeding: A Comparison of Models. *Crop Sci.*, *52*(1), 146–160. http://doi.org/10.2135/cropsci2011.06.0297

Iglesias, C., Mayer, J., Chavez, L., & Calle, F. (1997). Genetic potential and stability of carotene content in cassava roots. *Euphytica*, *94*(3), 367–373. http://doi.org/10.1023/A:1002962108315

Ikeogu, U. N., Davrieux, F., Dufour, D., Ceballos, H., Egesi, C. N., & Jannink, J.-L. (2017). Rapid analyses of dry matter content and carotenoids in fresh cassava roots using a portable visible and near infrared spectrometer (Vis/NIRS). *PLOS ONE*, *12*(12), e0188918. http://doi.org/10.1371/journal.pone.0188918

Iwata, H., Hayashi, T., Terakami, S., Takada, N., Sawamura, Y., & Yamamoto, T. (2013). Potential assessment of genome-wide association study and genomic selection in Japanese pear Pyrus pyrifolia. *Breeding Science*, *63*(1), 125–140. http://doi.org/10.1270/jsbbs.63.125

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

Jia, Y., & Jannink, J.-L. L. (2012). Multiple-Trait Genomic Selection Methods Increase Genetic Value Prediction Accuracy. *Genetics*, *192*(4), 1513–1522. http://doi.org/10.1534/genetics.112.144246

Kim, J. K., Lee, S. Y., Chu, S. M., Lim, S. H., Suh, S. C., Lee, Y. T., … Ha, S. H.

(2010). Variation and correlation analysis of flavonoids and carotenoids in Korean pigmented rice (Oryza sativa L.) cultivars. *Journal of Agricultural and Food Chemistry*, *58*(24), 12804–12809. http://doi.org/10.1021/jf103277g

Krinsky, N. I., & Johnson, E. J. (2005). Carotenoid actions and their relation to health and disease. *Molecular Aspects of Medicine*, *26*(6), 459–516. http://doi.org/10.1016/j.mam.2005.10.001

Lebot, V., Champagne, A., Malapa, R., & Shiley, D. (2009). NIR determination of major constituents in tropical root and tuber crop flours. *Journal of Agricultural and Food Chemistry*, *57*(22), 10539–10547. http://doi.org/10.1021/jf902675n

Lee, S., Choi, H., Cha, K., Kim, M. K., Kim, J. S., Youn, C. H., … Chung, H. (2012). Random forest as a non-parametric algorithm for near-infrared (NIR) spectroscopic discrimination for geographical origin of agricultural samples. *Bulletin of the Korean Chemical Society*, *33*(12), 4267–4270. http://doi.org/10.5012/bkcs.2012.33.12.4267

Lorenz, A. J., Chao, S., Asoro, F. G., Heffner, E. L., Hayashi, T., Iwata, H., … Jannink, J. L. (2011). *Genomic Selection in Plant Breeding. Knowledge and Prospects. Advances in Agronomy* (Vol. 110). http://doi.org/10.1016/B978-0-12-385531-2.00002-5

Ly, D., Hamblin, M., Rabbi, I., Melaku, G., Bakare, M., Gauch, H. G., … Jannink, J. L. (2013). Relatedness and genotype ?? environment interaction affect prediction accuracies in genomic selection: A study in cassava. *Crop Science*, *53*(4), 1312–1325. http://doi.org/10.2135/cropsci2012.11.0653

Maass, D., Arango, J., Wüst, F., Beyer, P., & Welsch, R. (2009). Carotenoid crystal formation in Arabidopsis and carrot roots caused by increased phytoene synthase protein levels. *PLoS ONE*, *4*(7). http://doi.org/10.1371/journal.pone.0006373

Meléndez-Martínez, A. J., Mapelli-Brahm, P., Benítez-González, A., & Stinco, C. M. (2015). A comprehensive review on the colorless carotenoids phytoene and phytofluene. *Archives of Biochemistry and Biophysics*. http://doi.org/10.1016/j.abb.2015.01.003

Metrohm. (2013). *NIR Spectroscopy : A guide to near-infrared spectroscopic analysis of industrial manufacturing processes*. *Monograph*. http://doi.org/8.108.5026EN – 2013-02

Möhring, J., & Piepho, H. P. (2009). Comparison of weighting in two-stage analysis of plant breeding trials. *Crop Science*, *49*(6), 1977–1988. http://doi.org/10.2135/cropsci2009.02.0083

Morillo C, Y., Sanchez, T., Morante, N., Ch�vez, A. L., Morillo C, A. C., Bola�os, A., & Ceballos, H. (2013). Preliminary study about the inheritance of the carotenoids content in roots from cassava (manihot esculenta crantz) segregating populations. *Acta Agronomica; Vol. 61, Nom. 3 (2012); 253-264 2323-0118 0120-2812*, *Volume 61*. Retrieved from https://translate.google.com/translate?hl=en&sl=es&u=https://revistas.unal.edu.co/index.php/acta_agronomica/article/view/37541/39917&prev=search

Mugode, L., Ha, B., Kaunda, A., Sikombe, T., Phiri, S., Mutale, R., … De Moura, F. F. (2014). Carotenoid retention of biofortified provitamin a maize (Zea mays L.) after Zambian traditional methods of milling, cooking and storage. *Journal of*

Agricultural and Food Chemistry, *62*(27), 6317–6325. http://doi.org/10.1021/jf501233f

Nweke, F. New challenges in the cassava transformation in Nigeria and Ghana (2004). Washington, D.C: International Food Policy Research Institute EPTD.

Okeke, U. G., Akdemir, D., Rabbi, I., Kulakow, P., & Jannink, J.-L. (2017). Accuracies of univariate and multivariate genomic prediction models in African cassava. *Genetics Selection Evolution*, *49*(1), 88. http://doi.org/10.1186/s12711-017-0361-y

Oliveira, E. J., Santana, F. A., Oliveira, L. A., & Santos, V. S. (2014). Genetic parameters and prediction of genotypic values for root quality traits in cassava using REML/BLUP. *Genetics and Molecular Research*, *13*(3), 6683–6700. http://doi.org/10.4238/2014.August.28.13

Owens, B. F., Gore, M. A., Magallanes-Lundback, M., Tiede, T., Diepenbrock, C. H., Kandianis, C. B., … Rocheford, T. (2014). A foundation for provitamin a biofortification of maize: Genome-wide association and genomic prediction models of carotenoid levels. *Genetics*, *198*(4), 1699–1716. http://doi.org/10.1534/genetics.114.169979

Paine, J. A., Shipton, C. A., Chaggar, S., Howells, R. M., Kennedy, M. J., Vernon, G., … Drake, R. (2005). Improving the nutritional value of Golden Rice through increased pro-vitamin A content. *Nature Biotechnology*, *23*(4), 482–487. http://doi.org/10.1038/nbt1082

Paiva, S. A., & Russell, R. (2013). β -Carotene and Other Carotenoids as Antioxidants Review Series : Antioxidants and their Clinical Applications. *Journal of the American College of Nutrition*, *18*(October 2014), 37–41. http://doi.org/10.1080/07315724.1999.10718880

Pérez-Rodríguez, P., Gianola, D., González-Camacho, J. M., Crossa, J., Manès, Y., & Dreisigacker, S. (2012). Comparison Between Linear and Non-parametric Regression Models for Genome-Enabled Prediction in Wheat. *G3:Genes|Genomes|Genetics*, *2*(12), 1595–1605. http://doi.org/10.1534/g3.112.003665

Pérez, J. C., Lenis, J. I., Calle, F., Morante, N., Sánchez, T., Debouck, D., & Ceballos, H. (2011). Genetic variability of root peel thickness and its influence in extractable starch from cassava (Manihot esculenta Crantz) roots. *Plant Breeding*, *130*(6), 688–693. http://doi.org/10.1111/j.1439-0523.2011.01873.x

Piepho, H. P., Möhring, J., Schulz-Streeck, T., & Ogutu, J. O. (2012). A stage-wise approach for the analysis of multi-environment trials. *Biometrical Journal*, *54*(6), 844–860. http://doi.org/10.1002/bimj.201100219

Pillay, K., Siwela, M., Derera, J., & Veldman, F. J. (2014). Provitamin A carotenoids in biofortified maize and their retention during processing and preparation of South African maize foods. *Journal of Food Science and Technology*, *51*(4), 634–644. http://doi.org/10.1007/s13197-011-0559-x

Qi, Y. (2012). Random forest for bioinformatics. In *Ensemble Machine Learning: Methods and ApplicatiOns* (pp. 307–323). http://doi.org/10.1007/9781441993267_10

R Core Team. (2017). R: A Language and Environment for Statistical Computing. *R*

Foundation for Statistical Computing. http://doi.org/http://www.R-project.org/
Rabbi, I., Hamblin, M., Gedil, M., Kulakow, P., Ferguson, M., Ikpan, A. S., … Jannink, J. L. (2014). Genetic mapping using genotyping-by-sequencing in the clonally propagated cassava. *Crop Science*, *54*(4), 1384–1396. http://doi.org/10.2135/cropsci2013.07.0482
Rabbi, I. Y., Hamblin, M. T., Kumar, P. L., Gedil, M. A., Ikpan, A. S., Jannink, J. L., … Zhu, F. (2013). Sustainability of cassava (Manihot esculenta Crantz) as industrial feedstock, energy and food crop in Nigeria. *Crop Science*, *53*(4), 1–8. http://doi.org/10.1017/CBO9781107415324.004
Rabbi, I. Y., Udoh, L. I., Wolfe, M., Parkes, E. Y., Gedil, M. A., Dixon, A., … Kulakow, P. (2017). Genome-Wide Association Mapping of Correlated Traits in Cassava: Dry Matter and Total Carotenoid Content. *The Plant Genome*, *10*(3). http://doi.org/10.3835/plantgenome2016.09.0094
Rodriguez-Amaya, D. B. D. ., & Kimura, M. (2004). HarvestPlus Handbook for Carotenoid Analysis. *HarvestPlus Technical Monographs*, 59. Retrieved from http://ebrary.ifpri.org/utils/getfile/collection/p15738coll2/id/125148/filename/125149.pdf
Roggo, Y., Chalus, P., Maurer, L., Lema-Martinez, C., Edmond, A., & Jent, N. (2007). A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. *Journal of Pharmaceutical and Biomedical Analysis*, *44*(3 SPEC. ISS.), 683–700. http://doi.org/10.1016/j.jpba.2007.03.023
Sánchez, T., Ceballos, H., Dufour, D., Ortiz, D., Morante, N., Calle, F., … Davrieux, F. (2014). Prediction of carotenoids, cyanide and dry matter contents in fresh cassava root using NIRS and Hunter color techniques. *Food Chemistry*, *151*, 444–451. http://doi.org/10.1016/j.foodchem.2013.11.081
Schulz-Streeck, T., Ogutu, J. O., & Piepho, H. P. (2013). Comparisons of single-stage and two-stage approaches to genomic selection. *Theoretical and Applied Genetics*, *126*(1), 69–82. http://doi.org/10.1007/s00122-012-1960-1
Shenk, J. S., & Westerhaus, M. O. (1991). Population Definition, Sample Selection, and Calibration Procedures for Near Infrared Reflectance Spectroscopy. *Crop Science*, *31*(2), 469. http://doi.org/10.2135/cropsci1991.0011183X003100020049x
Sila, A. M., Shepherd, K. D., & Pokhariyal, G. P. (2016). Evaluating the utility of mid-infrared spectral subspaces for predicting soil properties. *Chemometrics and Intelligent Laboratory Systems*, *153*, 92–105. http://doi.org/10.1016/j.chemolab.2016.02.013
Smith, A., Cullis, B., & Thompson, R. (2001). Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics*, *57*(4), 1138–1147. http://doi.org/10.1111/j.0006-341X.2001.01138.x
Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redoña, E., … McCouch, S. R. (2015). Genomic selection and association mapping in rice (Oryza sativa): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PLoS Genetics*, *11*(2), e1004982. http://doi.org/10.1371/journal.pgen.1004982

Strobel, M., Tinz, J., & Biesalski, H. K. (2007). The importance of β-carotene as a source of vitamin A with special regard to pregnant and breastfeeding women. *European Journal of Nutrition*. http://doi.org/10.1007/s00394-007-1001-z

Svetnik, V., Liaw, A., Tong, C., Christopher Culberson, J., Sheridan, R. P., & Feuston, B. P. (2003). Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences*, *43*(6), 1947–1958. http://doi.org/10.1021/ci034160g

Teye, E., Huang, X., & Afoakwa, N. (2013). Review on the Potential Use of Near Infrared Spectroscopy (NIRS) for the Measurement of Chemical Residues in Food. *American Journal of Food Science and Technology*, *1*(1), 1–8. http://doi.org/10.12691/ajfst-1-1-1

Tobias, R. D. (1995). An introduction to partial least squares regression. *SAS Conference Proceedings: SAS Users Group International 20 (SUGI 20)*, 2–5. http://doi.org/http://support.sas.com/techsup/technote/ts509.pdf

Tosato, M. G., Orallo, D. E., Fangio, M. F., Diz, V., Dicelio, L. E., & Churio, M. S. (2016). Nanomaterials and natural products for UV-photoprotection. In *Surface Chemistry of Nanobiomaterials* (pp. 359–392). http://doi.org/10.1016/B978-0-323-42861-3.00012-1

van Vliet, T., van Schaik, F., Schreurs, W. H., & van den Berg, H. (1996). In vitro measurement of beta-carotene cleavage activity: methodological considerations and the effect of other carotenoids on beta-carotene cleavage. *International Journal for Vitamin and Nutrition Research. Internationale Zeitschrift Für Vitamin- Und Ernährungsforschung. Journal International de Vitaminologie et de Nutrition*, *66*(1), 77–85. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8698551

Vinet, L., & Zhedanov, A. (2011). A 'missing' family of classical orthogonal polynomials. *Journal of Physics A: Mathematical and Theoretical*, *44*(8), 085201. http://doi.org/10.1088/1751-8113/44/8/085201

Welsch, R., Arango, J., Bär, C., Salazar, B., Al-Babili, S., Beltrán, J., … Beyer, P. (2010a). Provitamin A Accumulation in Cassava (Manihot esculenta) Roots Driven by a Single Nucleotide Polymorphism in a Phytoene Synthase Gene. *The Plant Cell*, *22*(10), 3348–3356. http://doi.org/10.1105/tpc.110.077560

Welsch, R., Arango, J., Bär, C., Salazar, B., Al-Babili, S., Beltrán, J., … Beyer, P. (2010b). Provitamin A Accumulation in Cassava (*Manihot esculenta*) Roots Driven by a Single Nucleotide Polymorphism in a Phytoene Synthase Gene. *The Plant Cell*, *22*(10), 3348–3356. http://doi.org/10.1105/tpc.110.077560

Wolc, A., Arango, J., Settar, P., Fulton, J. E., O'Sullivan, N. P., Preisinger, R., … Dekkers, J. C. M. (2011). Persistence of accuracy of genomic estimated breeding values over generations in layer chickens. *Genetics Selection Evolution*, *43*(1), 23. http://doi.org/10.1186/1297-9686-43-23

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. In *Chemometrics and Intelligent Laboratory Systems* (Vol. 58, pp. 109–130). Elsevier. http://doi.org/10.1016/S0169-7439(01)00155-1

Wolfe, M. D., Pino, D., Carpio, D., Alabi, O., Ezenwaka, L. C., Ikeogu, U. N., … Jannink, J.-L. (2017). Prospects for Genomic Selection in Cassava Breeding.

Plant Genome, *10*(3). http://doi.org/10.3835/plantgenome2017.03.0015
Wolfe, M. D., Rabbi, I. Y., Egesi, C., Hamblin, M., Kawuki, R., Kulakow, P., … Jannink, J. (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), 0. http://doi.org/10.3835/plantgenome2015.11.0118
Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, *88*(1), 76–82. http://doi.org/10.1016/j.ajhg.2010.11.011
Zhang, X., Pérez-Rodríguez, P., Semagn, K., Beyene, Y., Babu, R., López-Cruz, M. A., … Crossa, J. (2015). Genomic prediction in biparental tropical maize populations in water-stressed and well-watered environments using low-density and GBS SNPs. *Heredity*, *114*(3), 291–9. http://doi.org/10.1038/hdy.2014.99

CHAPTER 4: SNP-BASED PARENTAGE EVALUATION OF A POLYCROSS NURSERY IN THE GENOMIC SELECTION BREEDING SCHEME OF CASSAVA

Abstract

The use of controlled crossing scheme in cassava to obtain sufficient botanical seeds and ensure random crosses among selected progenitors is quite challenging. Breeders often rely on some form of open pollination to increase seed production, although such measures lack full parentage account. Within this era of abundant genomic information in cassava, we propose the use of polycross nursery system, a form of open-pollination involving the inter-mating of selected individuals in isolation from other compatible clones, to generate sufficient seeds and to use available high density markers in resolving the parentage dilemma associated with this scheme. We established polycross nurseries in two locations in two years and generated sufficient seeds (16000 and 25000, respectively) to drive meaningful selection in cassava. We used two methods - realized relationship and penalized regression, to assign parentage in unknown complete parents-progeny pair cases using SNP markers on both simulated and empirical datasets. In all the simulated cases, we measured the accuracy of correctly assigning parents to a given progeny compared to its recorded true parents and obtained 100% accuracy under various assumptions of simple versus complex inheritance, presence and absence of genotyping error etc. Using empirical records from the nursery scheme of breeding programs, accuracy of predicting both recorded parents ranged from 11% to 60%, 57% to 98% for predicting at least one of the parents, 22% to 76% for male and 33% to 80% in predicting the recorded female parents across both methods. We suggested likely issues that can affect expected nursery record. Besides abundant seeds, the possibility of random mating was high with less than 1% selfing rate using the assignment approaches in analysing one of the polycross scheme. Therefore, with good random design and flowering control, the genome-enabled polycross system promises to promote adequate random mating while increasing the chances of identifying desirable candidates in cassava. The parentage assignment methods used here are very fast and easy to implement, with no need for special genotype format or additional biological information. This setup can potentially enable the estimate of both specific (SCA) and general combining abilities (GCA) necessary for understanding heterotic patterns in cassava.

Keywords: Cassava, Genomic selection, Parentage assignment, Polycross.

Introduction

The full-sib controlled crosses is often ideal for the hybridization and genetic recombination of selected parents with desirable trait loci in many breeding programs. However, in cassava breeding the success rate of controlled crosses, in terms of seed

production and random combination of favorable loci from the candidate parents, is often hampered by many factors not limited to differences in flowering periods, where some clones flower as early as 1 to 2 months after planting (MAP) while others only flower after 6 to 8 MAP; differences in anthesis, even among clones with similar flowering periods; high abortion rate associated with physical handling; loss of viability of pollen during storage; and high cost and the lack of technical skill required in making crosses (Byrne 1984; Halsey et al. 2008; Ceballos et al. 2017). In addition, cassava has a trilocular fruit and from any successful cross, the average expected number of seeds per fruit is only about one or two and a maximum of three (Ceballos et al. 2004; Jennings 1963). Also, seed viability and germination rate could be very poor, further adding to the challenges of mating superior genotypes and overall decrease in the expected genetic improvement of cassava (Maria Gonçalves Fukuda, de Oliveira, and Iglesias 2002; Jennings and Iglesias 2002; Ndubuisi et al. 2015). As a result of these bottlenecks in generating adequate crosses for evaluation among selected clones, cassava breeders often resort to some form of open pollination in order to increase seed production for viable selection process. For example, the use of natural (insect-mediated) crosses was reported to be more successful than hand pollinations in an interspecific cross in cassava (NASSAR 1989).

The polycross nursery, a form of open pollination involving the isolation and natural inter-mating of selected clones apart from other compatible unselected clones (Nduwumuremyi, Tongoona, and Habimana 2013), has been occasionally utilized in cassava breeding (Mulualem and Bekeko 2015) and other crops as well (Riday, Smith, and Peel 2015; Tysdal and Crandall 1947; Aastveit and Aastveit 1990; J. Li, Jongsma,

and Wang 2014). The scheme is cheap, easy to implement and promotes random mating among selected genotypes (Tysdal and Crandall 1947; Jennings and Iglesias 2002). However, the use of polycross scheme could pose a challenge in breeding programs due to the lack of full parental information thereby, restricting the estimate of variance components and genetic gain (Nduwumuremyi, Tongoona, and Habimana 2013; Nguyen and Sleper 1983). However, there have been attempts in using molecular markers, mostly polymorphic microsatellites, to resolve the lack of complete parental record in evolution and behavioral ecology (Hughes 1998; Soares et al. 2006), animal (Van Eenennaam et al. 2007), and plant breeding programs (Aastveit and Aastveit 1990; Riday, Smith, and Peel 2015), including cassava (Vincent et al. 2014). But the statistical approaches available for parentage assignment, including the use of exclusive, fractional, categorical allocation and likelihood-based methods could be limited by genotyping errors (Jones et al. 2010; Christie 2013). Recently, a Bayesian approach, which accounts for genotyping errors and missing data, was introduced (Christie et al. 2013) but similar to other approaches, these parentage assignment methods are more effective with bi-allelic genotype formats, and reduced datasets, and sometimes additional information such as biological behavior and morphological information are needed for accurate analyses (Drábek 2009; Neff 2001).

With the increasing advancement in genotyping technology coupled with efficient computing techniques and computer capacity, abundant and informative single nucleotide polymorphic (SNP) markers are increasingly becoming available in many breeding programs, and there is an increasing interest in using SNP markers for parentage analyses (Van Eenennaam et al. 2007; Werner et al. 2004; Heaton et al. 2014).

Markers have been useful in estimating relationships or co-ancestry based on the fact that individuals sharing lots of genotypes at different SNPs are likely because they belong to the same family and SNP-based relationships have been argued to be more accurate than pedigree methods (Mucha, Wolc, and Strabel 2010; Medrano-E'Vers et al. 2017). Whereas progenies are assumed to receive a random half of each parent's genes and full-sibs are expected to share half their genes in pedigree methods, genomic data provide information on which half of each parent's genes an individual receives and precisely the proportion of genes shared by full-sibs (Medrano-E'Vers et al. 2017). The prospects of inferring parentage from pairwise relationships have been discussed extensively, and estimates of genealogical co-ancestry has been derived from molecular data (Glaubitz, Rhodes, and Dewoody 2003; Toro, García-Cortés, and Legarra 2011). Similarly, parentage assignment has been considered a regression problem using a constrained non-linear optimization in regressing the genotype of a given individual with unknown parentage against the genotypes of possible parents (Boerner 2017; Boerner and Banks 2016).

The benefits of accurate SNP marker paternity assignment enabling the use polycross scheme especially under the current genomic selection effort (Wolfe et al. 2017) are enormous. It will lead to drastic reduction in seed production cost and further shortens the long breeding cycle of cassava (Ceballos et al. 2004; Ceballos et al. 2015), while increasing the possibility of genetic recombination and the overall likelihood of identifying desirable individuals that are high yielding, widely adaptable and resistant to major diseases and pests. As a monoecious and outcrossing crop and with adequate field design, the chances of self-pollination in a polycross scheme in cassava is expected

to be very low (Da Silva, Bandel, and Martins 2003; Kawano et al. 1978). The staggering of planting dates has been recommended in addressing any known differences in flowering habit among candidate parents (Ceballos et al. 2004). Besides good estimation of genetic gain and provision of information for parentage evaluation (Acquaah 2012), the genome-enabled parentage assignment offers great opportunity for the estimate of specific (SCA) and general combining abilities (GCA). The analyses could provide useful insight on possible rate of selfing, differential pollen donor rate and possible evidence of incompatibility between pairs of selected parents in cassava. More importantly, accurate parentage evaluation is relevant in resolving cases of known technical mixtures and the verification of crosses.

In this study, we used information from a SNP-based kinship matrix and a constrained regression model - penalized LASSO (Least Absolute Shrinkage and Selection Operator) in assigning parents to a sample of individuals from a polycross scheme in the genomic selection pipeline in cassava. The two approaches are fast and easy to implement and do not require any modification of input data. They can be implemented using the same sets of markers used for other genomic studies.

Materials and methods

Polycross nursery and genotyping: Twenty nine (29) clones, a subset of selected clones with high genomic estimated breeding values (GEBVs) and selection index (SI) for certain selection traits from the genomic selection scheme at National Root Crops Research Institute (NRCRI), Umudike, Nigeria were used to establish a polycross nursery at two locations and two cropping seasons- Umudike in 2014/2015 (Poly1) and Ubiaja in 2015/2016 (Poly2). The selected clones were laid out in a randomized scheme

with 29 replications per clone using a quasi-complete Latin squares design (Bailey 1984) with some manually modified to discourage self-pollination by ensuring that replicates of the same clone were not closely situated. Available data of the selected parents on flowering (Appendix 4.1) was used to ensure that individuals with similar flowering rate and time were used as parents in the polycross scheme. An isolation of about 50 and 100 meters were observed to avoid pollen contamination from any neighboring cassava plot at Umudike and Ubiaja, respectively. An isolation of 30 meters had been previously suggested (Halsey et al. 2008; Kawano et al. 1978). At maturity, seeds from different replications of the 29 clones were harvested and processed.

The harvested seeds were planted out in a seedling nursery and leaf samples were collected for genotyping from a random selection of varying number of individuals per clone after being transplanted to the field. A total of 944 samples from Poly1 were genotyped from an average of 33 seeds (33 x 29) derived from 3 to 8 replications of the 29 parent clones (Figure 4.1). The collected leaf samples were processed at the NRCRI, Umudike biotechnology laboratory and shipped to the Genomic Diversity Facility at Cornell for genotyping- by- sequencing (GBS) (Elshire et al., 2011) with Apek-1 restriction enzyme (Hamblin and Rabbi 2014). The protocol used in processing the genotype data has been reported and included a combination of custom scripts and common variant call file manipulation tools (Wolfe et al. 2016). SNPs were called with TASSEL 5.0 GBS pipeline (version 2) (Glaubitz et al. 2014) and aligned to the cassava reference genome (version 6.1) (Bredeson et al. 2016). Genotypes were either called when a minimum of two reads was present or imputed with Beagle (version 4.0) (Browning and Browning 2008). Individuals with more that 80% missing data and

markers with more than 60% missing genotype calls were removed. Equally markers with extreme deviation from Hardy–Weinberg equilibrium ($X^2 > 20$) were removed (Wolfe et al. 2016).

Data sets:

Simulated datasets: Three datasets were simulated to reflect cases of simple and complex inheritance structures, different segregation patterns, crosses versus selfed as well as the presence of some degree of genotyping error. We modeled a case of uniform allele frequency on small data set involving one progeny from each parent-pair and no genotyping error (Sim I). The Sim I data consisted of 29 progenies generated from 29 bi-parents with 200 loci (SNP). Sim II was simulated using information from an empirical data set and to have multiple progenies per parent-pair with some degree of genotyping error (10%). The allele frequencies and the number of markers were derived from genotype data of the International Institute of Tropical Agriculture's (IITA) training population (C0). We simulated a total of 600 progenies from 60 parents, and 113246 SNP markers. Sim I and II were simulated using SOLOMON package (R Core Team 2017; Christie 2010). The underlying assumption in Sim I and II is that markers were independently and randomly passed from parents to progenies.

On the other hand, we considered a case of nonrandom segregation using founder haplotypes by simulating another dataset using 28 parents out of the 29 polycross parents with 112570 markers reflecting another empirical genotype file from NRCRI, Umudike (Sim III). The simulation was carried out using the "Breeding Scheme Language" package in R (Yabe, Iwata, and Jannink 2017). The program equally

provided opportunity for a random selfing of some of the parents. Therefore, Sim III contains both cross and self-pollinated progenies.

All the datasets were fully characterized, such that we have the full record of all the true parents and their offspring.

Empirical datasets: We used a subset of the IITA's nursery data (C0 and C1) with full pedigree record of parents-progeny pair to evaluate the accuracy of the two parentage assignment methods (Emp I). The data containing written record of parent-offspring information consisted total of 61 parents (41 female and 42 male parents), 730 progenies with varying number of progenies per parent-pair and 113246 SNP markers. Similarly, nursery data from the National Crops Resources Research Institute (NaCRRI), Uganda was used to validate the parentage methods (Emp II). The NaCCRI data consists of both bi-parental crosses and progenies harvested from open-pollination. The total number of parents involved in both pollination schemes was 80 with 384 progenies from bi-parental crosses, 764 progenies from open-pollination and 46760 common SNP markers between parents and progenies. We eventually used the two parentage assignment methods in analyzing the 2014/2015 polycross nursery from NRCRI (Poly1) – Emp III. The data were made up of 29 parents and 944 progenies with 78823 common SNP markers between the parents and progenies.

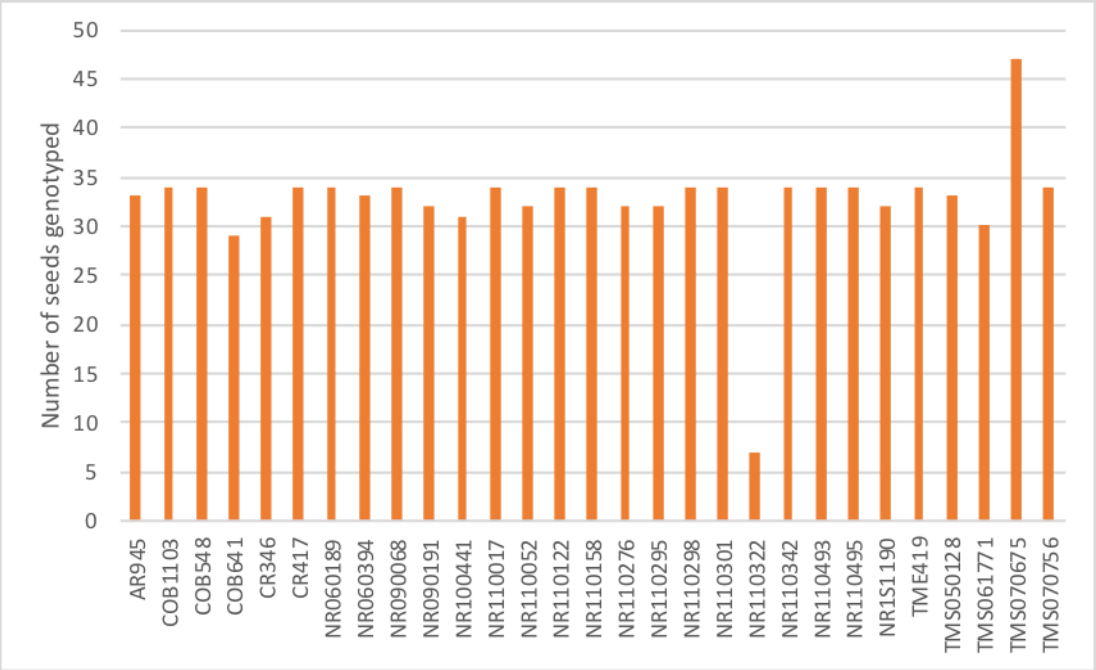

Figure 4.1: Number of progenies genotyped per clone from the written field record

Statistical analyses:

The two-parentage assignment including the realized relationship (RR) and penalized regression (PR) approaches were used to evaluate the simulated and empirical datasets. All analyses were carried out in R platform (R Core Team 2017).

Realized Relationship (RR) approach: We derived a genetic relationship between each individual progeny with incomplete parentage record and all the putative parents from a genetic covariance between each candidate progeny at a time and all the putative parents defined as: $G = \frac{WW'}{c}$ where $W_{ik} = X_{ik} + 1 - 2p_k$; p_k = the frequency of the 1 allele at marker k and c = normalization constant ($c = 2\sum_k P_k(1 - P_k)$) and X is a matrix (*n \times m*) of unphased genotypes for *n* lines and *m* biallelic markers (Endelman and Jannink 2012; VanRaden 2008). The relationship or kinship coefficient between each given progeny and the putative parents is defined as the probability that an allele

sampled at random from a progeny is identical to an allele sampled at random from the same locus in the parents due to descent (Falconer and Mackay 1996; Lacy 2012). The relationship matrix was derived using *A.mat* function in *rrBLUP* package (Coster 2017; Endelman and Jannink 2012). The pairwise values obtained from the relationship matrix (kinship coefficient) between the parents and each progeny were ranked and the two topmost parents with the highest kinship coefficients were nominated as the likely parents of that given progeny.

Penalized regression: *We used* LASSO as a penalized regression model in assigning parentage to unknown or incomplete parentage cases. LASSO is both a shrinkage and selection regression model and involves the addition of absolute penalty term to the objective function that causes the shrinkage of the regression coefficients - L1-norm:

$\hat{\beta}^{lasso} = argm\,||y - X\beta|_2^2 + \lambda\,\|\beta\|_1$ (Tibshirani 1996; Breheny 2015).

We used LASSO to regress the individual genotype of a given progeny with unknown parentage to that of all the potential parents one at a time:

$Y = X\beta + E$, where y = genotype of a given progeny; X = genotype of all the putative parents; β = coefficients of parental genotypes and E = Error.

The penalized regression was executed using *penalized* package in R (Goeman et al. 2017). The derived penalized regression coefficients were reported in percentage as the ratio of each observed coefficient to the sum of coefficients from the putative parents for any given progeny. We considered two scenarios representing two possible cases of complete loss of parent records and partial loss of information on one of the parents, which in most cases should be the male parent. The second scenario is a typical case in

the open-pollination of many breeding crops where seeds are harvested from a female plants of known identity making the incomplete parents record really the problem of loss of information of the pollen donor. We modeled the first scenario where there is a possibility of complete loss of parents' information by allowing penalty on all the potential parents (PR1) whereas in the second scenario, we allowed penalty on all the potential parents except the recorded female parents even though they were still part of the regression (PR2). As in the realized relationship method, we selected in each progeny case the two topmost parents with the highest regression coefficients as the likely parents of the given progeny.

Results and Discussion:

Polycross seeds: From the polycross nurseries established at Umudike (Poly1) and Ubiaja (Poly2), we harvested and processed over 16,000 and 48,000 seeds, respectively, excluding some unprocessed seeds from the two experiments. Seeds were collected and recorded from each clone and the number of seeds per clone processed in Poly1 ranged from 52 seeds in clone "NR061089" to 1263 seeds in "TMS061771" while in Ploy2, it ranged from 13 seeds in "NR110052" to 4863 seeds in "NR110495" (Figure 4.2; Appendix 4.2). There were cases of missing replications, especially in Poly1 (Table 4.1) and as such, the number of seeds per clone obtained from the two polycross experiments might not completely reflect the flowering or seed retention rate of the various clones. However, seed production per clone varied across the two locations and overall, the total number of seeds per clone obtained from Poly2 is about three times more than what was obtained in Poly1. In general, Ubiaja has been identified as having favorable

conditions for flowering in cassava, resulting in its use as a hybridization site for IITA and NRCRI (Adeyemo et al. 2017). Considering the number of clones and the number of seeds processed from the two experiments, the average number of seeds per clone was about 550 seeds in Poly1 and 1655 seeds in Poly2. With a high tendency of hybridization in a polycross set-up, the scheme increases the tendency of identifying progenies with desirable attributes from the potential random recombination of loci of the selected parents. A case of average number of progenies per parent used in driving genetic improvement in cassava was reported as 255 progenies obtained from both controlled and half-sib crosses over years (Ceballos et al. 2016). Therefore, the use of polycross scheme could facilitate abundant seeds in a single year rather than waiting for about two years to generate sufficient seeds before embarking on extensive evaluation and selection processes (Ceballos et al. 2016; Mulualem and Bekeko 2015). The scheme is cost effective and could reduce the cost of generating botanical seeds as well as the overall cost of breeding (Jennings and Iglesias 2002; Tysdal and Crandall 1947). The expected random crossing in all possible combinations among the selected progenitors makes it possible to estimate both specific and general combining ability (breeding values) of the progenitors. This could equally be useful in identifying families with heterotic benefits in cassava (Nguyen and Sleper 1983; Ceballos et al. 2016).

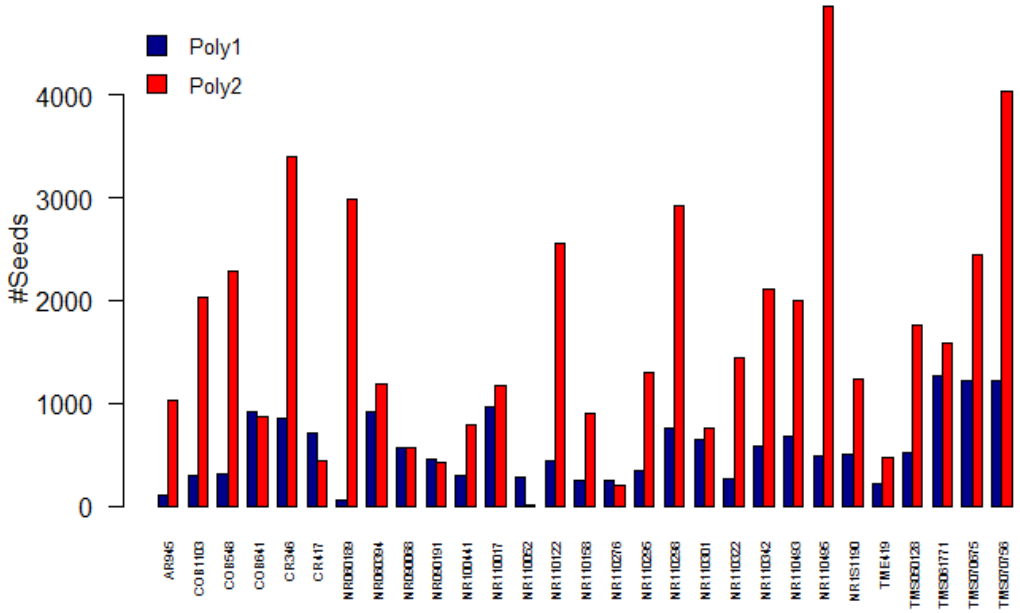

Figure 4.2: Number of seeds per clone generated from polycross scheme established at Umudike in 2014/2015 (Poly1) and Ubiaja in 2015/2016 (Poly2) cropping seasons.

Table 4.1: Germination and flowering evaluation of the polycross field at Umudike (Poly1) at 5MAP

	Number of plants	Percentage
Number of missing plants @ 5MAP	197	23.45%
Number of non-flowering @ 5MAP	497	59.17%
Number of flowering clones @ 5MAP	146	17.38%

Simulated data – simple and complex cases: In all the simulated cases (Sim I –III), the two approaches were able to correctly identify the recorded true parents of all the progenies in all the different cases (Table 4.2). The result of the penalized regression

method was consistent in both cases where there was penalty on all the parents (PR1) and only on a set of candidate male parents (PR2). Where there were selfed parents in Sim III, the two methods were able to accurately recognize the selfed parents as the topmost candidates with the highest regression or kinship coefficients in each progeny case.

Table 4.2: Accuracy of parentage assignment using realized relationship method and penalized regression (with and without condition on female parents) on simulated and empirical datasets.

Data	Method	Prediction Accuracy (%)			
		2P	>1P	M	F
Sim I, II and III	RR	100	100	-	-
	PR	100	100	-	-
Emp I	RR	49.12	98.22	70.78	76.54
	PR	59.81	96.16	76.13	79.84
Emp II_BC	RR	10.67	56.8	22.13	45.33
	PR	21.64	62.8	28.76	55.67
Emp II_OP	RR	-	-	-	68.72
	PR	-	-	-	66.67
Emp III	RR	-	-	-	34.96
	PR	-	-	-	33.26

In practical breeding, especially in cassava, there might be chances of both cross and self-pollination even though cassava is an outcrossing species (Halsey et al. 2008); it was therefore, imperative to differentiate between such instances. A further assessment

of the penalized regression and kinship coefficients especially using Sim III with known cases of both selfing and true crosses, we observed that the coefficients of the topmost predicted parents were usually higher in selfed cases than the coefficients obtained in true crosses in realized relationship (Figure 4.3; Appendix 4.3) and penalized regression (Figure 4.4; Appendix 4.4) methods. Furthermore, we considered the ratio of topmost parent's coefficients to that of the second topmost parents. The average ratio of 1.18 and a range of 1 to 2.19 for realized relationship and average of 1.11 and a range of 1 to 1.5 were observed in crossed cases. The selfed cases had a minimum of 3.06 and 18 from realized and penalized regression methods respectively. These criteria are perhaps very important in separating between likely incidences of selfing and true crosses in a nursery.

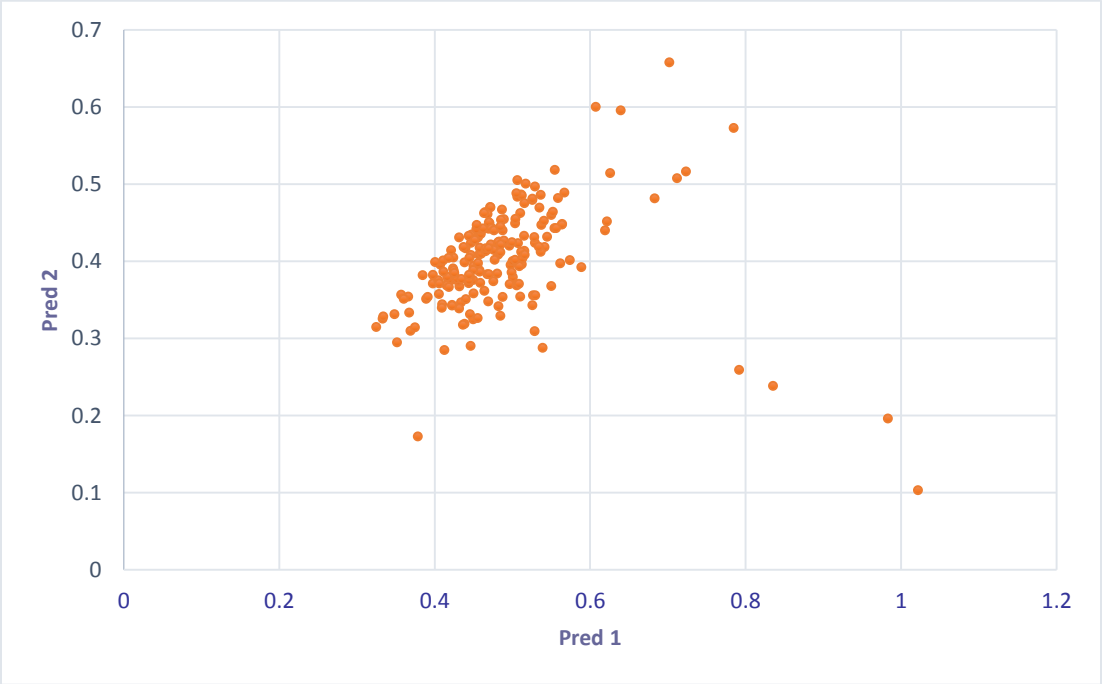

Figure 4.3: Relationship coefficients of the topmost (Pred 1) and second parents (Pred 2) using realized relationship method on Sim III – the last three points on x-axis with values above 0.8 were selfed.

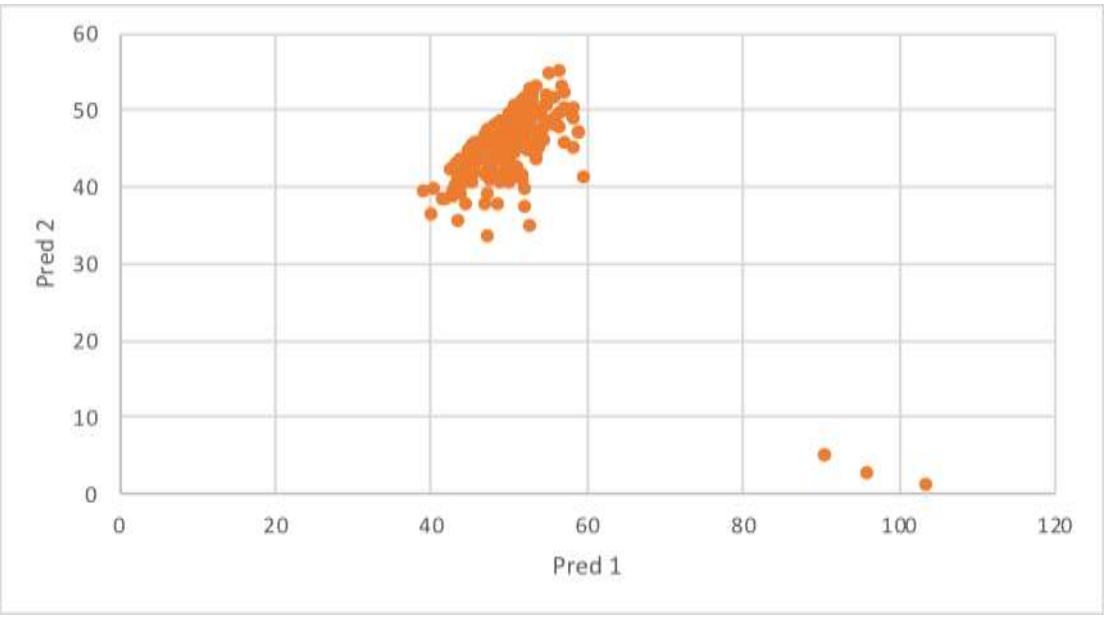

Figure 4.4: Regression coefficients of the topmost (Pred 1) and second parents (Pred 2) using penalized regression method on Sim III – the last three points on x-axis with values above 80 were selfed.

Empirical data: From IITA's crossing nursery data with full parent record (Emp 1), we used the two parentage methods to assign parents to the progenies and compared that with the written records. The analyses were carried out for the 730 C1 progenies obtained from bi-parental crosses among a subset of the C0 population. Using penalized regression method, we still considered the possibility of complete (PR1) as well as incomplete unknown parental records (PR2). We defined assignment accuracy (in percentage) as the number of occasions where the assigned top two parents from the two parentage methods matched the written crossing record over the total number of progenies evaluated. Accuracies were estimated where the two topmost predicted parents matched the two written parents (2P), at least one of the parents (1P) as well as predicting either the written male (M) or female (F) parents. Similarly, accuracies were

estimated and compared against the written records from NaCCRI in both the bi-parental crosses (Emp II_BC) and open-pollination (Emp II_OP) cases.

The result of the analyses showed various degrees of deviation from the written parentage across the breeding programs using the two parentage methods in 2P, 1P, M and F cases (Table 4.2). Comparable to the simulated results, the two forms of penalized regression (PR1 and PR2) gave the same accuracies and therefore, the reported accuracies from penalized regression method are from the two conditional options of penalty.

Generally, the penalized regression method had higher prediction accuracy than the realized relationship approach, except in the case of predicting at least one of the recorded parents (>1P) in Emp I, where the realized method was higher (98%) than the penalized approach (96%) and the cases of open pollination in Emp II_OP and Emp III (Table 4.2). From Emp I and II_BC, the result of the evaluation showed that the accuracy of assignment for both parents (2P) using penalized regression approach was about 11% more than using the realized relationship method. Otherwise, in predicting at least one of the parents in Emp II_BC, only the male and only the female cases, the penalized method was slightly higher than the realized relationship method in Emp I and Emp II_BC. In all the single parent predictions (male and female), prediction accuracies were higher in female than male parents using both penalized and realized relationship methods (Table 4.2).

The two topmost predicted parents were compared to the written female record from the NRCRI polycross (Emp III) using both the realized relationship and penalized regression methods. The accuracies of assignment were approximately 35% using

realized relationship and 33% from penalized regression methods. The low accuracies were expected following reports of row and column orientation mix-up during planting and the harvesting of seeds.

Examining the number of clones that were represented among the two topmost assigned parents among all the selected parents from the NRCRI polycross, the entire 29 parents appeared as one of the top two parents using RR while 28 parents appeared in PR with the number of occurrences ranging from 1 to 230 depending on the method (Figure 4.5). This is a good indication of the importance of the scheme in promoting random mating and sustaining optimum effective population size in cassava crossing nursery. The low number of events for some clones could be traceable to mislabeling of clones that was mentioned earlier. This was evidenced on the low prediction accuracy of the female clones where seeds were harvested. The expected number of events per clone/family was supposed to be higher than the number of progenies sampled for genotyping from each clone. However, the high number of clones represented among the topmost two parents from the two assignment methods is an indication of the possibility of using polycross to promote random mating among selected parents even though only about 6% of the realized seeds were genotyped and used for the analyses. We used the top two parents since any of the clones could have served as the male of female plant.

When using the criteria identified earlier in separating between crossed and selfed-pollinated cases, that is 0.8 or 80% coefficients as a threshold in RR and PR respectively, it was observed that very few individuals (about 7 individuals, less than one percent of the total genotyped progenies) had coefficients greater than 80% from the PR method (Figure 4.6) and realized relationship (Figure not presented). In essence, self-pollination

might not be a serious issue in the polycross scheme following a good randomization and added that cassava is naturally an out-crossing crop (Halsey et al. 2008; Hernán Ceballos, Hershey, and Becerra-López-Lavalle 2012).

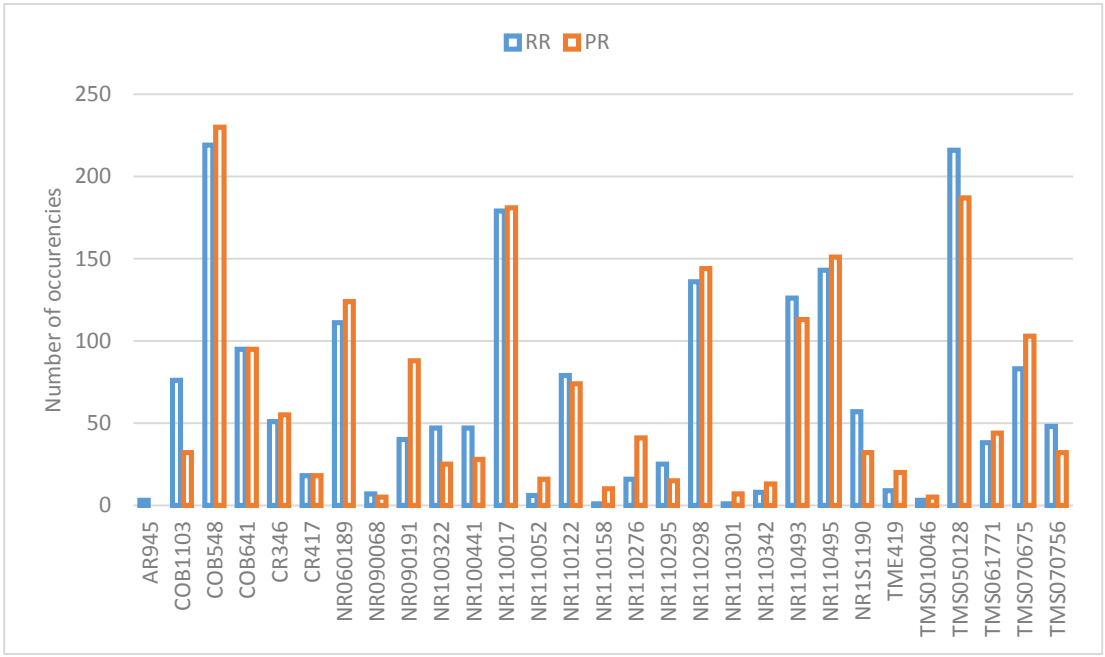

Figure 4.5: Number of occurrences of clones as the predicted top two parents using realized relationship (RR) and penalized regression (PR) methods.

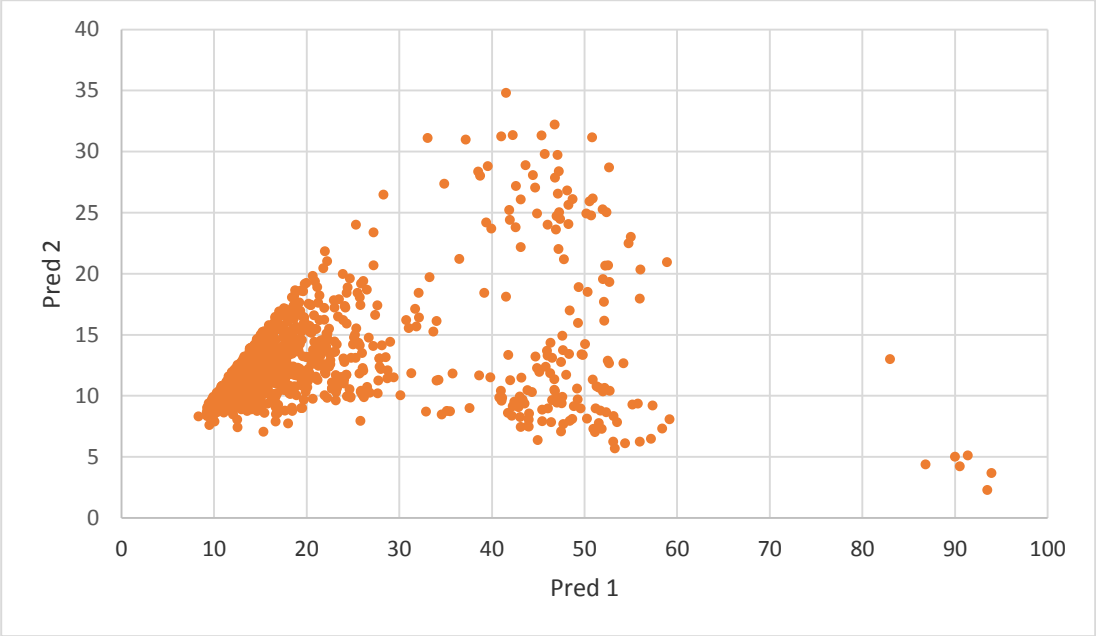

Figure 4.6: Regression coefficients of the topmost parents (Pred 1) and the runner-up parents (Pred 2) using penalized regression method on Emp III.

In practical cassava breeding just as in other crops, ideally except in cases of clone mislabeling, the uncertainty of parentage identity should be less of a female parent's issue than the pollen donor especially in wind and insect pollinated crops, since seeds are still usually attached and harvested from the female parent. Mislabeling of clones could take the form of an outright misplaced identity where different seeds/clones/stakes are planted and labelled differently from the intended seeds/clones/stakes of interest or the switching of male and female records. Therefore, there should be serious emphasis on quality control during various pre-planting, planting and post-planting events such as stem cutting, transportation, stakes preparation, planting to avoid contaminations. Clone misidentification could arise from regenerated volunteer clones especially if the field was used previously for cassava production.

On the other hand, pollen mixtures could potentially bias the outcome of an intended crossing scheme in a breeding program (Vincent et al. 2014; Pennisi 2006; Sk Lai et al. 2010). Although cassava pollen grains are relatively large in size compared to other crops, the grains are still small enough (90 to 150μm in diameter) and sticky (Halsey et al. 2008) to adhere to pollination tools during crosses. This could potentially promote pollen contamination.

The impact of pollination bags on seed contamination has been reported in few other crops (McAdam, Senior, and Hayward 1987; Schaffert, Virk, and Senior 2016) but need to be fully investigated in cassava. The underlying consideration in using pollination bags during cassava pollination should include the need to ward-off insects from direct contact with flowers and more importantly, the ability of the bags to prevent unwanted pollen from reaching the flowers through any potential pores. It is important to consider the effect of wear and tear arising from bag re-usage (within and across seasons) and the general condition of crossing bags on the genetic purity of crosses in cassava. The highlighted factors among others could distort the certainty of a written nursery record.

We therefore, recommend proper labelling in crossing nurseries and more importantly, in polycross scheme where many replicates of a considerable number of clones need to be carefully laid-out in a well-designed random format. Current adoption of modern breeding tools such as the use of barcodes will help to address certain drawbacks associated with field mislabeling and misplacement of planting materials.

Also, better understanding of the genetics of flowering will be complementary in the establishment and management of crossing scheme including polycross nursery in

cassava (Bull et al. 2017; Adeyemo et al. 2017). Where available, flowering records will help to facilitate good field design and nursery establishment. Understanding and adoption of best flower manipulation options including the use of hormones, grafting and day length controls (Ceballos et al. 2017; Adeyemo et al. 2017; Bai et al. 2017) will improve flowering conditions in cassava. Planting could be done multiple times to account for possible differences in flowering period although various flowering groups usually overlap (Hernán Ceballos et al. 2004; Halsey et al. 2008).

Important environmental and soil factors affecting flowering are vital for a successful polycross set-up (Keating 1982; Ravi and Ravindran 2006; Haukka, Dreyer, and Esler 2013). The previous flowering record on the clones used in establishing the polycross scheme especially at Umudike (Poly1) had similar flowering time and branching habit, which is highly associated with flowering (Adeyemo et al. 2017; Ceballos et al. 2017) (Appendix 4.1), however, the flowering evaluation carried out the next year at 5MAP showed that only 17.4% of the clones flowered early (Table 4.1). There were about 23% missing stands in the scheme which was later traced to the soil condition of the nursery site. The soil was reported to be hard, clayey and prone to termite infestation which affected the germination of some clones; however, that was the available area with such intended level of isolation from neighboring cassava fields in the institute at that time. We had a better growth, flowering and germination conditions at Ubiaja in Poly2 which had been known to have favorable climatic and soil conditions for robust flowering and seed production (Adeyemo et al. 2017). This could have reflected in the seed output from the two locations (Figure 4.2; Appendix 4.2). In addition, the

polycross in Ubiaja was situated in the middle of a yam breeding field away from any cassava plots.

Improvement in genotyping protocols and processing could enhance better results in parentage assignment. The genotype files used in this study for the parents and progenies were processed in different years.

Parental assignment in a polycross scheme has obvious benefits and the proposed assignment methods could be valuable in parental verification and possible parental reconstructions in a general crossing nursery scheme in cassava. As the cost and quality of genotyping keep improving and supporting the adoption of new breeding techniques such as GS (Wolfe et al. 2017; Eder Jorge de Oliveira et al. 2012), the use of high density SNP markers will continue to be an effective tool in cassava improvement. The scheme promises to promote adequate inter-mating in all possible combinations as well as the number of seeds per bi-parental cross of selected parents. This has a direct implication in increasing the effective population and improving the selection intensity in cassava which in turn will help promote gain in this crop (Campo and Turrado 1997; Lorenz et al. 2011). As reported earlier, with a sample of about 6% from Poly1, the number of parents that showed up as potential parents from the two assignment methods suggest the tendency of using the polycross scheme in supporting random mating in cassava. Notwithstanding, the various degrees of occurrences per clone calls for caution to ensure that few parents with robust pollen potential do not dominate the scheme even though this could be addressed by proper layout, the use of good planting materials as well as siting the experiment in a favorable location that enhances adequate flowering of clones.

In addition to adequate seed generation, the proposed complementary schemes could substantially reduce pollination cost and labor (Jennings and Iglesias 2002). Under the GS pipeline, there is likely no added genotyping cost associated with polycross scheme as the genotyping cost would have been factored into the GS cost.

The two proposed methods are straightforward and easy to implement with the same set of markers used for other genomic studies. There is no limit to the number of loci required for analysis and there is no need to reformat the genotype data (Jones et al. 2010; Christie 2010; Christie et al. 2013; Jones and Ardren 2003). In terms of run time, the methods were computationally efficient (Table 4.3). In general, the realized relationship method had faster execution time than the penalized regression options. Under the penalized options, the case of penalizing only the male parents (PR2) saved some time when compared to PR1 where all the parents were penalized (Table 4.3).

Table 4.3: System run time (in seconds) using the relationship and penalized methods on Sim I, Sim II and IITA's data with a 3.4 GHz, 32GB iMac computer.

Data	Method	User	System	Elapsed
Sim I	RR	0.182	0.013	0.193
	PR1	0.563	0.042	0.593
	PR2	0.352	0.041	0.582
Sim II	RR	2030.002	302.156	2259.029
	PR1	26602.457	4557.442	25756.746
	PR2	17620.630	4282.594	18464.466
Sim III	RR	963.079	48.992	270.620
	PR2	1632.147	124.558	204.406
Emp I	RR	3164.955	623.492	662.807
	PR1	4268.879	1142.276	1816.446
	RR	3008.249	623.522	482.782
	PR1	5214.061	1257.177	2195.906
Emp III	RR	2137.523	196.243	2302.177
	PR1	9505.236	2092.775	10161.679
	PR2	6361.722	1379.535	6939.488

Conclusion

The high level of accuracy from the different practical conditions from the simulated datasets gives good confidence in using the proposed methods in parentage assignment and resolving the lack of full parentage record in cassava polycross scheme. Even though the accuracies of simulated data were higher than the empirical data, the methods gave high accuracies (96% and 98% using PR and RR respectively) in predicting either of the two written parents using IITA's data (Emp II). There is therefore need for improved field and technical control in cassava crossing nurseries. The percentage of progenies sampled for analyses from Poly1 was about 6% and might be insufficient to fully estimate the rate of pollen donor per clone. Also, the number of seeds per clone recorded might not necessarily be used to evaluate the seed production potential of the clones since so many seeds were not processed in the field. However, further investigations and studies would be important to further improve the scheme. It would be important to understand the dynamics of insect population in the polycross scheme as well as the general cassava nursery system.

Many similar studies have focused on finding the optimum number of markers required for accurate parentage assignment (Boerner and Banks 2016; Christie 2010), we focused on solving an imminent problem in cassava breeding and made use of the abundant number of markers available to us. However, the outcome of this study is still relevant in many other breeding programs. The analyses were based on the assumption that all the progenitors (the true parents) are within the polycross scheme as we controlled against external pollination from unwanted pollen donors.

With accurate field design and control of many factors related to cassava flowering, it is possible to use a polycross scheme for high seed production and random inter-mating of selected parents in cassava. Under the current GS scheme, the polycross system with full parental construction using high density markers promises to further shorten the breeding cycle of cassava while accelerating the accuracy and precision of identifying candidates with desirable traits. Compared to other methods of parental evaluation, the methods presented are fast, do not need a special genotype format and have no limits to the number of loci to be used. There are also no additional behavioral and morphological information required to run the models.

Acknowledgments

We acknowledge the efforts of all the breeding and hybridization teams at NRCRI and IITA for data collection. We also acknowledge members of Jannink/Sorrells' Lab. for their input.

APPENDICES

Appendix 4.1: Flowering data of some of the selected polycross parents.

S/N	Clones	F_rate	F_time	Plt_Ht	Br_Ht	Level_Br	No_Br
1	CR346	3	1	101.5	38	4	3
2	TMS050128	3	1	58.5	27	2	3
3	TMS070756	3	1	95	26.5	3	3
4	NR060394	3	1	126.5	37.5	4	3
5	TMS070675	3	1	141	18	5	4
6	AR945	3	1	146.5	38.5	5	3
7	COB1103	3	1	117.5	28	3	3
8	NR110122	3	1	111.5	39.5	4	3
9	TMS061771	3	1	104	16	4	4
10	CR417	3	1	118.5	30	5	3
11	COB64	3	1	111	43.5	4	3
12	NR1S1190	3	1	81	29.5	3	3
13	NR110342	3	1	111.5	39	3	3
14	NR110495	3	1	128	58	3	3
15	NR110158	3	1	139	17	5	3
16	NR060189	3	1	96.5	28	3	3
17	NR110017	3	1	127	48	3	3
18	NR100322	3	1	104.5	22.5	3	3
19	NR110295	3	1	105.5	41	3	3
20	NR090068	3	1	100	27.5	4	3
21	AR919	3	1	92.5	26.5	4	3
22	NR090191	3	1	117	25	4	4
23	NR110298	3	1	133.5	22	4	3
24	NR100441	3	1	103	36	3	3
25	NR110301	3	1	111.5	42.5	3	3
26	TMS010046	3	1	99	33	3	3
27	NR110052	3	1	120	78	3	3
28	NR110276	3	1	96.5	29.5	3	3
29	COB548	3	1	124.5	31.5	5	3
30	NR110493	3	1	134	39.5	3	3

F_rate = Flowering rate (1= poor, 3 = moderate and 5 = profuse); F_time = Flowering time (1= early: <4MAP, 3 = medium: 4 – 6MAP, 5 = late: > 6MAP); Plt_Ht = Plant height (cm); Br_Ht = Height at first branching (cm); Level_Br = Level of branching (1=poor, 3 moderate, 5 = highly branching) and No_Br = Number of branches.

Appendix 4.2: Number of seeds per clone processed from Poly1 and Poly2.

Clone	Poly1	Poly2
AR945	109	1026
COB1103	289	2034
COB548	307	2288
COB641	918	868
CR346	845	3396
CR417	713	435
NR060189	53	2982
NR060394	920	1193
NR090068	559	561
NR090191	454	422
NR100441	299	787
NR110017	963	1169
NR110052	279	13
NR110122	433	2555
NR110158	252	908
NR110276	248	203
NR110295	347	1292
NR110298	752	2931
NR110301	650	761
NR110322	260	1439
NR110342	583	2113
NR110493	680	1999
NR110495	494	4863
NR1S1190	506	1229
TME419	208	478
TMS050128	514	1757
TMS061771	1263	1580
TMS070675	1216	2452
TMS070756	1216	4036

Appendix 4.3: Summary of realized relationship coefficients on all the datasets.

	Pred1			Pred2			Ratio (Pred1/Pred2)		
Dataset	Min	Max	Av	Min	Max	Av	Min	Max	Av
Sim I	0.29	0.47	0.37	0.26	0.39	0.32	1	1.3	1.15
Sim II	0.33	0.34	0.33	0.29	0.31	0.3	1.07	1.15	1.11
Sim III_C	0.32	0.78	0.48	0.17	0.66	0.41	1	2.19	1.18
Sim III_S	0.79	1.02	0.91	0.1	0.26	0.2	3.06	9.93	5.38
Emp I_2	0.14	0.38	0.22	0.1	0.33	0.18	1	2.1	1.23
Emp I_1	0.1	0.48	0.24	0.06	0.48	0.24	1	2.77	1.13
Emp I_0	0.08	0.44	0.23	0.07	0.26	0.18	1	3.23	1.35
Emp II_{BC}	0.03	0.7	0.17	0.03	0.38	0.12	1	4.23	1.45
Emp II_{SC}	0.04	0.72	0.18	0.04	0.47	0.12	1	3.69	1.52

Sim I = Simulated data 1; Sim II = Simulated data 2; Sim III_C = Simulated data 3, cross-pollinated; Sim III_S = Simulated data 3, self-pollinated; Emp I_2 = IITA data where predicted parents matched the written records; Emp I_1 = IITA data where predicted parents matched at least one of the written records; Emp I_0 = IITA data where predicted parents matched none of the written records; Emp II_{BC} = NaCCRI data, cross-pollinated and Emp II_{SC} = NaCCRI data, self-pollinated.

Appendix 4.4: Summary of penalized regression coefficients on all the datasets.

	Pred1			Pred2			Ratio (Pred1/Pred2)		
Dataset	Min	Max	Av	Min	Max	Av	Min	Max	Av
Sim I	32.84	43.15	38.18	30.67	40.95	34.68	1	1.4	1.1
Sim II	45.34	47.06	46.13	38.26	40.07	39.19	1.1	1.2	1.18
Sim III_C	39.26	59.65	50.1	33.31	54.87	45.45	1	1.5	1.11
Sim III_S	90.49	103.62	95.15	1.29	4.92	3.4	18	80	39.01
Emp I_2	24.34	49.35	38.97	17.6	42.93	31.68	1	2.22	1.28
Emp I_1	12.52	74.84	37.36	8.19	44.38	25.56	1	6.08	1.53
Emp I_0	13.61	76.45	32.36	5.77	37.23	21.64	1	13.25	1.85
Emp II_{BC}	6.05	75.31	33.02	3.3	39.07	15.82	1	22.45	2.88
Emp II_{SC}	5.42	80.42	37.83	3.02	39.32	12.89	1	23.42	4.17

Sim I = Simulated data 1; Sim II = Simulated data 2; Sim III_C = Simulated data 3, cross-pollinated; Sim III_S = Simulated data 3, self-pollinated; Emp I_2 = IITA data where predicted parents matched the written records; Emp I_1 = IITA data where predicted parents matched at least one of the written records; Emp I_0 = IITA data where predicted parents matched none of the written records; Emp II_{BC} = NaCCRI data, cross-pollinated and Emp II_{SC} = NaCCRI data, self-pollinated.

REFERENCES

Aastveit, A. H., & Aastveit, K. (1990). Theory and application of open-pollination and polycross in forage grass breeding. *Theoretical and Applied Genetics*, *79*(5), 618–624. http://doi.org/10.1007/BF00226874

Acquaah, G. (2012). *Principles of Plant Genetics and Breeding: Second Edition*. *Principles of Plant Genetics and Breeding: Second Edition*. http://doi.org/10.1002/9781118313718

Adeyemo, O. S., Chavarriaga, P., Tohme, J., Fregene, M., Davis, J., & Setter, T. L. (2017). Overexpression of Arabidopsis FLOWERING LOCUS T (FT) gene improves floral development in cassava (Manihot esculenta , Crantz). *PLoS ONE*, 1–15. http://doi.org/10.1371/journal.pone.0181460

Bai, S., Tuan, P. A., Saito, T., Ito, A., Ubi, B. E., Ban, Y., & Moriguchi, T. (2017). Repression of TERMINAL FLOWER1 primarily mediates floral induction in pear (Pyrus pyrifolia Nakai) concomitant with change in gene expression of plant hormone-related genes and transcription factors. *Journal of Experimental Botany*, *68*(17), 4899–4914. http://doi.org/10.1093/jxb/erx296

Bailey, R. A. (1984). Quasi-complete Latin squares: construction and randomization. *J. R. Stat. Soc., Ser. B*, *46*(2), 323–334. Retrieved from http://www.jstor.org/stable/2345518

Boerner, V. (2017). On marker-based parentage verification via non-linear optimization. *Genetics Selection Evolution*, *49*(1), 50. http://doi.org/10.1186/s12711-017-0324-3

Boerner, V., & Banks, R. (2016). SNP based parentage verification via constraint non-linear optimisation. *Interbull Bulletin*, *0*(50). Retrieved from https://journal.interbull.org/index.php/ib/article/view/1395

Bredeson, J. V, Lyons, J. B., Prochnik, S. E., Wu, G. A., Ha, C. M., Edsinger-Gonzales, E., … Rokhsar, D. S. (2016). Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nature Biotechnology*, *34*(5), 562–570. http://doi.org/10.1038/nbt.3535

Breheny, P. (2015). The group exponential lasso for bi-level variable selection. *Biometrics*, *71*(3), 731–740. http://doi.org/10.1111/biom.12300

Browning, B. L., & Browning, S. R. (2008). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American Journal of Human Genetics*, *84*(2), 210–223. http://doi.org/10.1016/j.ajhg.2009.01.005

Bull, S., Alder, A., Barsan, C., Kohler, M., Hennig, L., Gruissem, W., & Vanderschuren, H. (2017). FLOWERING LOCUS T Triggers Early and Fertile Flowering in Glasshouse Cassava (Manihot esculenta Crantz). *Plants*, *6*(2), 22. http://doi.org/10.3390/plants6020022

Byrne, D. (1984). Breeding cassava. In *Plant Breeding Reviews* (Vol. 2, p. 73–134.). Hoboken, NJ, USA: John Wiley & Sons, Inc. http://doi.org/10.1002/9781118060995.ch3

Campo, J. L., & Turrado, H. (1997). Population size and selection intensity effects on short-term response for a selection index in Tribolium. *Journal of Animal Breeding and Genetics*, *114*(1–6), 107–119. http://doi.org/10.1111/j.1439-

0388.1997.tb00498.x
Ceballos, H., Hershey, C., & Becerra-López-Lavalle, L. A. (2012). New Approaches to Cassava Breeding. *Plant Breeding Reviews*, *36*, 427–504. http://doi.org/10.1002/9781118358566.ch6
Ceballos, H., Iglesias, C. A., Pérez, J. C., & Dixon, A. G. O. (2004). Cassava breeding: Opportunities and challenges. *Plant Molecular Biology*, *56*(4), 503–516. http://doi.org/10.1007/s11103-004-5010-5
Ceballos, H., Jaramillo, J. J., Salazar, S., Pineda, L. M., Calle, F., & Setter, T. (2017). Induction of flowering in cassava through grafting, *9*(February), 19–29. http://doi.org/10.5897/JPBCS2016.0617
Ceballos, H., Kawuki, R. S., Gracen, V. E., Yencho, G. C., & Hershey, C. H. (2015). Conventional breeding, marker-assisted selection, genomic selection and inbreeding in clonally propagated crops: a case study for cassava. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *128*(9), 1647–67. http://doi.org/10.1007/s00122-015-2555-4
Ceballos, H., Pérez, J. C., Barandica, O. J., Lenis, J. I., Morante, N., Calle, F., … Hershey, C. H. (2016). Cassava breeding I: The value of breeding value. *Frontiers in Plant Science*, *7*(AUG2016), 1–12. http://doi.org/10.3389/fpls.2016.01227
Christie, M. R. (2010). Parentage in natural populations: Novel methods to detect parent-offspring pairs in large data sets. *Molecular Ecology Resources*, *10*(1), 115–128. http://doi.org/10.1111/j.1755-0998.2009.02687.x
Christie, M. R. (2013, December). Bayesian parentage analysis reliably controls the number of false assignments in natural populations. *Molecular Ecology*. http://doi.org/10.1111/mec.12528
Christie, M. R., Tennessen, J. A., Blouin, M. S., & Barrett, J. (2013). Genetics and population analysis Bayesian parentage analysis with systematic accountability of genotyping error, missing data and false matching, *29*(6), 725–732. http://doi.org/10.1093/bioinformatics/btt039
Coster, A. (n.d.). Pedigree Functions [R package kinship2 version 1.6.4]. Retrieved from https://cran.r-project.org/web/packages/pedigree/index.html
Da Silva, R. M., Bandel, G., & Martins, P. S. (2003). Mating system in an experimental garden composed of cassava (Manihot esculenta Crantz) ethnovarieties. *Euphytica*, *134*(2), 127–135. http://doi.org/10.1023/B:EUPH.0000003644.60126.4a
de Oliveira, E. J., de Resende, M. D. V., da Silva Santos, V., Ferreira, C. F., Oliveira, G. A. F., da Silva, M. S., … Aguilar-Vildoso, C. I. (2012). Genome-wide selection in cassava. *Euphytica*, *187*(2), 263–276. http://doi.org/10.1007/s10681-012-0722-0
Drábek, J. (2009). Use of animal/plant freeware for calculating likelihood ratio for paternity and kinship in complicated human pedigrees. *Forensic Science International: Genetics Supplement Series*, *2*(1), 469–471. http://doi.org/10.1016/j.fsigss.2009.08.195
Endelman, J. B., & Jannink, J.-L. (2012). Shrinkage Estimation of the Realized Relationship Matrix. *G3:Genes|Genomes|Genetics*, *2*(11), 1405–1413.

http://doi.org/10.1534/g3.112.004259
FALCONER, D. S., & MACKAY, T. F. . (1996). Introduction to quantitative genetics., 463. http://doi.org/10.1002/bimj.19620040211
Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., & Buckler, E. S. (2014). TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. *PLoS ONE*, *9*(2), e90346. http://doi.org/10.1371/journal.pone.0090346
Glaubitz, J. C., Rhodes, O. E., & Dewoody, J. A. (2003). Prospects for inferring pairwise relationships with single nucleotide polymorphisms. *Molecular Ecology*, *12*(4), 1039–1047. http://doi.org/10.1046/j.1365-294X.2003.01790.x
Goeman, A. J., Meijer, R., Chaturvedi, N., Lueder, M., Rcpp, I., & Rcpp, L. (2017). Package ‘ penalized ,’ *1*. Retrieved from https://cran.r-project.org/web/packages/penalized/penalized.pdf
Halsey, M. E., Olsen, K. M., Taylor, N. J., & Chavarriaga-Aguirre, P. (2008). Reproductive biology of cassava (Manihot esculenta Crantz) and isolation of experimental field trials. *Crop Science*. http://doi.org/10.2135/cropsci2007.05.0279
Hamblin, M. T., & Rabbi, I. Y. (2014). The effects of restriction-enzyme choice on properties of genotyping-by-sequencing libraries: A study in Cassava (Manihot esculenta). *Crop Science*. http://doi.org/10.2135/cropsci2014.02.0160
Haukka, A. K., Dreyer, L. L., & Esler, K. J. (2013). Effect of soil type and climatic conditions on the growth and flowering phenology of three Oxalis species in the Western Cape, South Africa. *South African Journal of Botany*, *88*, 152–163. http://doi.org/10.1016/j.sajb.2013.07.012
Hayes, B. J., Panozzo, J., Walker, C. K., Choy, A. L., Kant, S., Wong, D., … Spangenberg, G. C. (2017). Accelerating wheat breeding for end-use quality with multi-trait genomic predictions incorporating near infrared and nuclear magnetic resonance-derived phenotypes. *Theoretical and Applied Genetics*, *1*(0123456789). http://doi.org/10.1007/s00122-017-2972-7
Heaton, M. P., Leymaster, K. A., Kalbfleisch, T. S., Kijas, J. W., Clarke, S. M., McEwan, J., … Chitko-Mckown, C. G. (2014). SNPs for parentage testing and traceability in globally diverse breeds of sheep. *PLoS ONE*, *9*(4). http://doi.org/10.1371/journal.pone.0094851
Hughes, C. (1998, March). Integrating molecular techniques with field methods in studies of social behavior: A revolution results. *Ecology*. Ecological Society of America. http://doi.org/10.1890/0012-9658(1998)079[0383:IMTWFM]2.0.CO;2
Jennings, D. L. (1963). Variation in pollen and ovule fertility in varieties of cassava, and the effect of interspecific crossing on fertility. *Euphytica*, *12*(1), 69–76. http://doi.org/10.1007/BF00033595
Jennings, D. L., & Iglesias, C. (2002). *Cassava: biology, production and utilization*. *Cassava: biology, production and utilization*. Wallingford: CABI. http://doi.org/10.1079/9780851995243.0000
Jones, A. G., & Ardren, W. R. (2003, October). Methods of parentage analysis in natural populations. *Molecular Ecology*. Blackwell Science Ltd. http://doi.org/10.1046/j.1365-294X.2003.01928.x

Jones, A. G., Small, C. M., Paczolt, K. A., & Ratterman, N. L. (2010). A practical guide to methods of parentage analysis. *Molecular Ecology Resources*. http://doi.org/10.1111/j.1755-0998.2009.02778.x

Kawano, K., Amaya, A., Daza, P., & Rios, M. (1978). Factors Affecting Efficiency of Hybridization and Selection in Cassava1. *Crop Science*, *18*(3), 373. http://doi.org/10.2135/cropsci1978.0011183X001800030005x

Keating, B. a. J. P. E. and S. F. (1982). Environmental Effects on Growth and Development of Cassava (Manihot Esculenta Crantz.) I. Crop Development. *Field Crops Research*, *5*, 271–281. Retrieved from https://ac.els-cdn.com/0378429082900302/1-s2.0-0378429082900302-main.pdf?_tid=a1ea59b0-2905-417f-bf14-8cc1778cf375&acdnat=1522968661_a7f55a7e9e34b7a19389cfe45ffb4b3b

Lacy, R. C. (2012). Extending pedigree analysis for uncertain parentage and diverse breeding systems. *Journal of Heredity*, *103*(2), 197–205. http://doi.org/10.1093/jhered/esr135

Li, J., Jongsma, M. A., & Wang, C. Y. (2014). Comparative analysis of pyrethrin content improvement by mass selection, family selection and polycross in pyrethrum [Tanacetum cinerariifolium (Trevir.) Sch.Bip.] populations. *Industrial Crops and Products*, *53*, 268–273. http://doi.org/10.1016/j.indcrop.2013.12.023

Lorenz, A. J., Chao, S., Asoro, F. G., Heffner, E. L., Hayashi, T., Iwata, H., … Jannink, J. L. (2011). *Genomic Selection in Plant Breeding. Knowledge and Prospects. Advances in Agronomy* (Vol. 110). http://doi.org/10.1016/B978-0-12-385531-2.00002-5

Maria Gonçalves Fukuda, W., de Oliveira, S., & Iglesias, C. (2002). 2002, Brazilian Society of Plant Breeding. *Crop Breeding and Applied Biotechnology*, (3), 355–360. Retrieved from http://www.sbmp.org.br/cbab/siscbab/uploads/c8128f42-57b1-acce.pdf

McAdam, N. J., Senior, J., & Hayward, M. D. (1987). Testing the Efficiency of Pollination Bag Materials. *Plant Breeding*, *98*(2), 178–180. http://doi.org/10.1111/j.1439-0523.1987.tb01113.x

Medrano-E'Vers, A., Morales-Hernández, A. E., Valencia-López, R., & Hernández-Salcedo, D. R. (2017). Enfermedad granulomatosa crónica. *Medicina Interna de Mexico*, *33*(3), 407–414. http://doi.org/10.1007/s13398-014-0173-7.2

Mucha, S., Wolc, A., & Strabel, T. (2010). Comparison of methods for estimation of genetic covariance matrix from SNP or pedigree data utilised to predict breeding value. *BMC Proceedings*, *4*, S7. http://doi.org/10.1186/1753-6561-4-s1-s7

Mulualem, T., & Bekeko, Z. (2015). Assessment of conventional breeding on cassava and its physiological adaptive mechanisms: Implication for moisture stress. *Asian Journal of Agricultural Research*, *9*(2), 38–54. http://doi.org/10.3923/ajar.2015.38.54

NASSAR, N. M. A. (1989). BROADENING THE GENETIC BASE OF CASSAVA, Manihot esculenta Crantz, BY INTERSPECIFIC HYBRIDIZATION. *Canadian Journal of Plant Science*, *69*(3), 1071–1073. http://doi.org/10.4141/cjps89-129

Ndubuisi, D., Nathaniel, U., Ewa, F., & Egesi, C. (2015). Crossability and germinability potentials of some cassava (Manihot esculenta Crantz) progenitors

for selection. *Journal of Plant Breeding and Crop Science*, *7*(March), 61–66. http://doi.org/10.5897/JPBCS2014.0479

Nduwumuremyi, A., Tongoona, P., & Habimana, S. (2013). Mating Designs: Helpful Tool for Quantitative Plant Breeding Analysis. *J. Plant Breed. Genet*, *01*(03), 117–129. Retrieved from http://www.escijournals.net/JPBG

Neff, B. D. (2001). Genetic paternity analysis and bredding success in bluegill sunfish (Lepomis macrochiros). *Journal of Heredity*, *92*, 111–119. Retrieved from https://oup.silverchair-cdn.com/oup/backfile/Content_public/Journal/jhered/92/2/10.1093/jhered/92.2.111/2/111.pdf?Expires=1500586328&Signature=NRSpazt8CSoiPEF3cyt5wQoB2RE4bWsbSz7~Je3phZAoJ2pjI7DFY-306iO~ITbrOgUmzF7~FurPDWJwC2HL57gnHh9hrTP48ME5UcM-tyFuWwc~

Nguyen, H. T., & Sleper, D. A. (1983). Theory and application of half-sib matings in forage grass breeding. *Theoretical and Applied Genetics*, *64*(3), 187–196. http://doi.org/10.1007/BF00303763

Pennisi, E. (2006). Genetics. Pollen contamination may explain controversial inheritance. *Science*, *313*(5795), 1864. http://doi.org/10.1126/science.313.5795.1864

R Core Team. (2017). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*. http://doi.org/http://www.R-project.org/

Ravi, V., & Ravindran, C. S. (2006). Effect of soil drought and climate on flowering and fruit set in cassava (Manihot esculenta Crantz). *Advances in Horticultural Science*, *20*(2), 147–150. Retrieved from https://www.jstor.org/stable/42882473?seq=1#page_scan_tab_contents

Riday, H., Smith, M. A., & Peel, M. D. (2015). A simple model for pollen-parent fecundity distributions in bee-pollinated forage legume polycrosses. *Theoretical and Applied Genetics*, *128*(9), 1865–1879. http://doi.org/10.1007/s00122-015-2553-6

Schaffert, R. E. ., Virk, D. S., & Senior, H. (2016). Comparing pollination control bag types for sorghum seed harvest. *Journal of Plant Breeding and Crop Science*, *8*(8), 126–137. http://doi.org/10.5897/JPBCS2016.0580

Sk Lai, B., Funda, T., Liewlaksaneeyanawin, C., Klápště, J., Niejenhuis, A., Cook, C., … El-Kassaby, Y. A. (2010). Pollination dynamics in a Douglas-fir seed orchard as revealed by pedigree reconstruction. *Annals of Forest Science*, *67*(8), 808–808. http://doi.org/10.1051/forest/2010044

Soares, T. N., Telles, M. P. C., Resende, L. V, Silveira, L., Jácomo, A. T. A., Morato, R. G., … Brondani, C. (2006). Paternity testing and behavioral ecology: A case study of jaguars (Panthera onca) in Emas National Park, Central Brazil. *Genetics and Molecular Biology*, *29*(4), 735–740. http://doi.org/10.1590/S1415-47572006000400025

Tibshirani, R. (1996). Regression Selection and Shrinkage via the Lasso. *Journal of the Royal Statistical Society B*. http://doi.org/10.2307/2346178

Toro, M. Á., García-Cortés, L. A., & Legarra, A. (2011). A note on the rationale for estimating genealogical coancestry from molecular markers. *Genetics Selection Evolution*, *43*(1). http://doi.org/10.1186/1297-9686-43-27

Tysdal, H. M., & Crandall, B. H. (1947). The Polycross Progeny Performance as an Index of the Combining Ability of Alfalfa Clones1, *40*, 3. Retrieved from https://dl.sciencesocieties.org/publications/aj/pdfs/40/4/AJ0400040293/

Van Eenennaam, A. L., Weaber, R. L., Drake, D. J., Penedo, M. C. T., Quaas, R. L., Garrick, D. J., & Pollak, E. J. (2007). DNA-based paternity analysis and genetic evaluation in a large, commercial cattle ranch setting. *Journal of Animal Science*, *85*(12), 3159–3169. http://doi.org/10.2527/jas.2007-0284

VanRaden, P. M. (2008). Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science*, *91*(11), 4414–4423. http://doi.org/10.3168/jds.2007-0980

Vincent, K., Robert, K., Morag, F., Tadeo, K., Yona, B., & Peter, V. (2014). Identification of F 1 Cassava (Manihot esculenta Crantz) Progeny Using Microsatellite Markers and Capillary Electrophoresis. *American Journal of Plant Sciences*, *5*, 119–125. http://doi.org/10.4236/ajps.2014.51015

Werner, F. A. O., Durstewitz, G., Habermann, F. A., Thaller, G., Krämer, W., Kollers, S., … Fries, R. (2004). Detection and characterization of SNPs useful for identity control and parentage testing in major European dairy breeds. *Animal Genetics*, *35*(1), 44–49. http://doi.org/10.1046/j.1365-2052.2003.01071.x

Wolfe, M. D., Pino, D., Carpio, D., Alabi, O., Ezenwaka, L. C., Ikeogu, U. N., … Jannink, J.-L. (2017). Prospects for Genomic Selection in Cassava Breeding. *Plant Genome*, *10*(3). http://doi.org/10.3835/plantgenome2017.03.0015

Wolfe, M. D., Rabbi, I. Y., Egesi, C., Hamblin, M., Kawuki, R., Kulakow, P., … Jannink, J. (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), 0. http://doi.org/10.3835/plantgenome2015.11.0118

Yabe, S., Iwata, H., & Jannink, J. L. (2017). A simple package to script and simulate breeding schemes: The breeding scheme language. *Crop Science*, *57*(3), 1347–1354. http://doi.org/10.2135/cropsci2016.06.0538

CHAPTER 5: RECOMMENDATIONS

Harmonizing genetic diversity in cassava:

Cassava is a very important crop that provides staple food for about 800 million people world-wide (FAO, 2013). Over the years, cassava has remained poorly funded in terms of research and its production has been exclusively relegated to low-income, and smallholder farmers with minimum input and mechanization (Ceballos, Iglesias, Pérez, & Dixon, 2004; Cock, 1987). Conversely, the research trend in cassava over the past decade has taken a new turn which is traceable to its potential not only as a food security crop but as feed and a source of raw material for biofuels and many other industries, including pulp, textile, food etc. While there is an urgent need for a sustainable research commitment from the governments of various countries in sub-Sharan Africa (SSA), most of the current research funding is from non-governmental agencies outside the SSA region. The specific nature of funding and complete lack of adequate funding in some instances, especially in national breeding programs, contribute to restricting the research scopes to only a handful of activities with limited or narrow genetic resources. The opportunities provided by the current era of increased funding and collaboration should be harnessed to develop less fragmented global research activities in cassava. Compared to other crops such as maize and rice, cassava breeders need a better understanding of the global and regional sources of variation necessary for developing broad-base sustainable products. In the SSA, the available germplasm is likely limited due to the early and current interventions in curtailing disease outbreaks, mostly cassava mosaic disease (CMD) and cassava brown streak disease (CBSD), with little or no variation for quality and nutritional traits. On the other hand, most South American

germplasm is rich in nutritional qualities but lacks resistance to most of the biotic stresses prevalent in Africa. Differences in trait correlations especially between DMC and carotenoids have also been reported between the two regions, while greater yield potentials tend to be reported in the Asian region. It is therefore pertinent to intensify research efforts that will deepen the understanding of global variation for many complex traits and maximize regional germplasm diversity to enable the development of sustainable products with broad agronomic and quality traits. Attention should be given to locally grown materials across the globe in order to understand end-user preferences and why certain materials have persisted over the years. Coordination to broaden the genetic base is equally necessary within institutions and local regions, especially in the SSA, where specific parallel programs have persisted as different pools of germplasm were moved to different areas and in different periods for different purposes.

Understanding regional genetic diversities relative to the global panel will help clarify the pedigrees of local and cultivated germplasms being used in different regions and exploit diverse allelic variations for improving complex quantitative traits in cassava.

Increasing the recombination of alleles:

Plant breeding generally relies on the probability of recovering good genotypes from a favorable cross, and determining the optimum number of crosses as well as the number of progenies per cross have been an important part of the discussion (Huehn, 1996; Witcombe & Virk, 2001). In the case of rare recombination events, large linkage blocks or even entire chromosomes may be inherited unchanged from one parent or the other in a progeny. In cassava breeding, there is little information about the optimum number of crosses and progeny per cross that are necessary to make adequate progress in

identifying favorable genotypes, considering the heterozygous nature and likely recombination events in the crop. Early reports showed that three outstanding genotypes emerged from about 372 000 genotypes derived from 4130 crosses over a period of fourteen years from the CIAT/Asia breeding scheme (Ceballos et al., 2004; Kawano, 2003; Kawano et al., 1998). Similarly, a combined selection for resistance to CMD and bacterial blight in IITA usually begins with about 100 000 seedlings and reduced to about 3000 genotypes after the initial first stage screening at the seedling nursery (Ceballos et al., 2004; Kawano, 2003). With the advent of new breeding tools and selection for many important traits, there should be a well-designed strategy to maximize the chances of selecting genotypes with desirable traits and this will largely rely on developing schemes that will enable adequate recombination of alleles from desirable donor parents. This is especially important since the expected number of seeds from a successful cross in cassava is about 1 or 2. The available crossing record used in this study derived from separate breeding programs – IITA and NaCCRI (Emp I and II, respectively), highlights the need for increased crosses and increased numbers of progenies from a given parental pair. About 61 and 80 unique parents from Emp I and II respectively, were used in developing bi-parental population of 730 and 384 progenies, respectively and this ultimately shrinks the probability of adequate recombination of parents and selection of desirable genotypes.

The greater the number of parents involved in a cross, the lower the chances of increasing the number of progenies per cross. It is possible to use the available genomic markers for a judicious selection of few unique parents for recombination while increasing the possibility of greater number of progenies per cross. Also, research effort

needs to focus on the possibility of predicting the likelihood of a cross giving rise to favorable segregants.

The use of polycross has been recommended for increasing the number of seeds from a crossing nursery with increased tendency of random recombination among selected parents. Further research and tools should be developed in adopting it, not only in bi-parental crosses of multiple parents but also for selfing purposes. As the need to purge deleterious alleles in cassava intensifies, it is important to begin to think of efficient means of achieving that since the artificial generation of seeds could be challenging and time consuming.

Also, though the cost of genotyping continues to go down, it is still not cheap enough to genotype large number of populations for screening at the early stage of breeding, especially in low resource breeding programs. Screening strategies for large numbers of progenies should be given attention. Adequate recombination of favorable alleles will promote better understanding of genetic correlations among important economic traits in cassava and give room for other genomic studies.

Understanding end-user preferences and increasing small holder income capacity:

While many new genetic materials are being generated in various breeding programs, there is need to strengthen or where possible, develop a strong farmer participatory research or create indices for variety adoption by farmers in order to avoid the rejection and loss of research resources used for the development and release of new varieties. There is an obvious regional, cultural and gender end-user differences in cassava (Teeken et al., 2018) and it is important not only to identify these preferences but also

devise means of incorporating such preferences in selecting for donor parents and subsequent variety development. Apart from quality traits, end-user or small-holder farmers' agricultural practices should be embraced in the selection and development of new varieties. Although it is not an easy task but the reality is that a farmer might likely reject a variety that does not fit perfectly into his/her cropping practices. For instance, small-holder farmers usually intercrop cassava with other crops and often interplant different varieties within a unit such as mounds or ridges. Besides quality and yield potentials, the ability of the new varieties to fit into the farmers' cultural practices is something that is not usually given considerable attention. Among other traits, the ability to store long in the ground could be a likely uncommon preference desired by subsistence farmers. Therefore, beyond understanding of obvious small-holder and end-user preferences, it might be important to put into consideration or model the overall agronomic and cultural practices including weeding patterns, planting methods, etc. that can possibly affect the adoption of new varieties by farmers. Alternatively, there should be a constant structure to continue to educate and persuade especially small holder farmers to adopt best practices that will help them maximize yield. More research communication and extension are needed in solving this problem.

More so, as the demand for cassava products continues to increase on the global stage (Jansson C, Westerbergh A, Zhang J, Hu X, 2009; Ye, Li, Lin, & Zhan, 2017), small-holder farmers should be encouraged to take advantage of these emerging markets to earn more money and improve their livelihood. In addition to improved research and investment for increased production, the creation of sustainable market chain is very important in lifting small-holder farmers from poverty. Low profitability has been

discouraging production and lack of interest among low income families to engage in agriculture. Cassava provides different opportunities that low income farmers can maximize to make profits. While so many market barriers could affect profitability, breeding strategies could be used to address such challenges related to post-harvest deterioration and targeting breeding goals that will enable farmers take advantage of the high demand for starch, ethanol, dry chips for animal feed etc.

Diversifying research investment in cassava:

As research opportunities for genetic improvement of cassava continue to increase, it is equally important to diversify investment to include efficient tools that will help realize the full benefits of the genetic improvement strategies. Emphasis on rapid and standardized phenotyping for large-scale population screening has been emphasized. Other efficient and cheap technologies should be pursued to ensure for example, the planting of uniform stakes, good weeding methods, efficient harvesting and post-harvest related operations etc. In this study, in order to achieve adequate homogeneity of cassava roots, we fabricated a flexible grinder that rely on a portable generator set for power supply which provide an opportunity for field evaluations. More investments in barcode systems and applications for data capture and management etc., will complement the ongoing genomic studies.

Sustaining Capacity development and entrepreneurship:

Thankfully, many infrastructural and human capacity developments have been targeted as important cassava improvement goals over the past decades. There is a need to sustain the current trend especially in the SSA where most of the investments are supported by external grants. Greater commitment and political will from the government of various

countries in the SSA is highly important. Meanwhile, strategic management could be adopted to maximize the available resources in the different breeding programs across the region and globally. Adequate consideration on long-term relevance should be paramount in infrastructural investments and where possible, cost-effective collaborations should be pursued.

As human capacity is very important in sustaining current research efforts, young scientists and personnel, especially women, should be groomed to avoid a knowledge gap. Constant retooling to keep pace with new technologies will be necessary to improve their capacity. Similarly, field and laboratory technicians should be trained and reskilled to help them keep up with the developing trends in their respective fields. Collaborations should be expanded to include various public and private institutions with differs expertise across the globe.

While employment opportunities could be challenging in many settings, it is becoming very important to embrace entrepreneurship especially among younger scientists. The anticipated agricultural development in the SSA can only be possible through a well-coordinated synergy and interventions from both the public and private sectors. Entrepreneurship, business and leadership skills should be part of the training of the younger generation of scientists. The engagement of young people in entrepreneurship will help to curb unemployment and encourage investments from the private sectors. Incentives as well as low interest loans could be provided to young scientist and small-holder farmers interested in business to establish cassava and agricultural related outfits. This will significantly improve the livelihood of these farmers since cassava has the potential not only to provide food but to generate income and diversify economies as

well. As more political commitment is anticipated, demand-driven strategies and sustainable management of investments should be encouraged.

REFERENCES

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

Cock, J. H. (1987). Cassav: new potential for a neglected crop. *Field Crops Research*, *15*, 389–390. Retrieved from http://ciat-library.ciat.cgiar.org/Articulos_Ciat/Digital/SB211.C3_C5_C.2_Cassava_New_potential_for_a_neglected_crop.pdf

FAO. (2013). *Save and Grow: Cassava*. Retrieved from http://www.fao.org/3/a-i3278e.pdf

Huehn, M. (1996). Optimum number of crosses and progeny per cross in breeding self-fertilizing crops. I. General approach and first numerical results. *Euphytica*, *91*(3), 365–374. http://doi.org/10.1007/BF00033099

Jansson C, Westerbergh A, Zhang J, Hu X, S. C. (2009). Cassava,a potential biofuel crop in the People's Republic of China. *Appl Energy*, *86*(595–599). Retrieved from https://ac-els-cdn-com.proxy.library.cornell.edu/S0306261909002049/1-s2.0-S0306261909002049-main.pdf?_tid=40efbcbf-e6d5-477b-a26b-92d54e8ecc9d&acdnat=1530745177_3fdeb1e5fd9ca7e808aba0aed0a7165f

Kawano, K. (2003). Thirty Years of Cassava Breeding for Productivity—Biological and Social Factors for Success. *Crop Science*, *43*(4), 1325. http://doi.org/10.2135/cropsci2003.1325

Kawano, K., Narintaraporn, K., Narintaraporn, P., Sarakarn, S., Limsila, A., Limsila, J., … Watananonta, W. (1998). Yield improvement in a multistage breeding program for cassava. *Crop Science*, *38*(2), 325–332. http://doi.org/10.2135/cropsci1998.0011183X003800020007x

Teeken, B., Olaosebikan, O., Haleegoah, J., Oladejo, E., Madu, T., Bello, A., … Tufan, H. A. (2018). Cassava Trait Preferences of Men and Women Farmers in Nigeria: Implications for Breeding. http://doi.org/10.1007/s12231-018-9421-7

Witcombe, J. R., & Virk, D. S. (2001). Number of crosses and population size for participatory and classical plant breeding. In *Euphytica* (Vol. 122, pp. 451–462). http://doi.org/10.1023/A:1017524122821

Ye, F., Li, Y., Lin, Q., & Zhan, Y. (2017). Modeling of China's cassava-based bioethanol supply chain operation and coordination. *Energy*, *120*, 217–228. http://doi.org/10.1016/j.energy.2016.12.114