GENOMIC PREDICTIONS WITH INCLUSION OF ENVIRONMENTAL COVARIATES TO IMPROVE CASSAVA FOR DISEASE RESISTANCE AND YIELD

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
Alfred Adebo Ozimati
May 2019

GENOMIC PREDICTIONS WITH INCLUSION OF ENVIRONMENTAL COVARIATES TO IMPROVE CASSAVA FOR DISEASE RESISTANCE AND YIELD

Alfred Adebo Ozimati Ph.D.

Cornell University 2019

Cassava is the fourth largest source of calories in developing countries after, maize, rice and wheat. However, yield losses due to viral diseases such as cassava mosaic disease (CMD) and cassava brown streak disease (CBSD) continue to impact the production of cassava in Asia and Africa. In Sub-Saharan Africa, CBSD is considered more destructive particularly in the East, Central, and Southern parts of Africa. One of the major obstacles in breeding cassava for traits of preference such as fresh root yield and disease resistance is the long breeding cycle (8 to 10 years). Genomic selection (GS), which uses genome-wide DNA markers and phenotypic records from the training population (TP), could help shorten the cycle by enabling estimation of the breeding values (GEBVs) and total genetic value for selection candidates without phenotyping.

The National Crops Resources Research Institute (NaCRRI), in Uganda is among the first cassava breeding programs to implement genomic selection. The present study covers three main areas. First, we assessed the impact of accelerated breeding on genetic variation, level of inbreeding, and trait correlations after one cycle of GS. Second, we tested genomic prediction accuracies for agronomic and disease traits in light of genotype-by-environment (G x E) interactions, providing opportunities when breeding for a wide adaptation. In the third objective, we tested genomic prediction accuracies for CBSD-related traits across breeding program (predictions of CBSD resistance in W. African clones, where the disease is non-existent) as a pre-emptive breeding strategy.

The highlights of these three studies were that (i) there was genetic progress made for most traits from GS cycle zero (C_0) to cycle one (C_1). The results indicated that selection based on GEBVs did not erode the original genetic diversity of lines bred under a GS enabled breeding system. Based on these results, we do not expect GS to cause rapid inbreeding as clones are advanced from cycle to cycle (ii) Inclusion of G x E information in genomic prediction showed moderate to high prediction accuracies for CBSD-related traits plus other agronomic traits such as harvest index (HI), under the different cross-validation prediction schemes. However, the predictive ability for root and shoot weight per plot were generally lower across GS prediction models evaluated, except for a scenario of predicting unobserved environments. This result implies that selection can be made accurately for CBSD, dry matter content (DMC), and HI based on genomic prediction models that incorporated G x E estimates. However, additional phenotypic information may be needed for the clones, when also selecting for fresh root yield (iii) Moderate prediction accuracies were recorded for CBSD in West African clones for foliar disease symptom expression, but low prediction accuracies were observed for root necrosis. Based on these results, building a training set comprising West African clones is recommended to predict CBSD resistance in West Africa germplasm where the disease is yet non-existent.

The collective output of these interrelated studies serve as vital information to breeders for enabling inter-regional genomic prediction and reducing multi-environment trial costs, without compromising genetic diversity levels across generations. The implementation of genomics-assisted breeding has the potential to help substantially improve cassava production in the developing world.

BIOGRAPHICAL SKETCH

Alfred Ozimati was born in a small town called Arua, Uganda. At the age of 13, he went to a Seminary School for Secondary education. Here he discovered his passion for Science, and as such he pursued a Bachelor degree in Horticulture at Makerere University, Uganda. After completion, he proceeded to work as an Agricultural Credit Officer in a local bank for a year. With unsettled passion for science, Alfred left the bank and joined Makerere University to pursue a Master of Science in Plant Breeding and Seed System. He was among the first cohort to win a scholarship from Alliance for Green Revolution in Africa (AGRA) for a Master of Science degree.

Immediately after completion of his Master's degree in 2012, Alfred joined the National Cassava Breeding Program in Uganda as a Breeder. In 2013, he received a fellowship award from the NextGen Cassava Breeding Project, funded by the Bill and Melinda Gates Foundation to pursue a Ph.D. in Plant Breeding and Genetics at Cornell University, USA. He hopes to be a dedicated agricultural scientist with a long-term goal of contributing to increased food production through the application of innovations in plant breeding and genetics, particularly in an African context.

ACKNOWLEDGMENTS

I sincerely want to thank my major advisor **Dr. Jean-Luc Jannink**, who guided me selflessly from the initial ideas of the research projects to completion of my study. Similar thanks goes to my Ph.D. committee, **Dr. Michael Gore** and **Dr. Stewart Grey** for their contribution throughout the study period. In a special memory, I would like to remember the late **Dr. Martha Hamblin**, who received us with a warm and motherly heart to Ithaca. Forever, you will be remembered for your kind heart. I also want to thank the leadership of National Crops Resources Research Institute (NaCRRI), especially **Dr. Yona Baguma**, the initial Investigator for the NextGen project at NaCRRI, who ensured that funds were available to conduct all the field experiments. Last, but not least I want to sincerely thank **Dr. Robert Kawuki**, the cassava breeder at NaCRRI for his mentorship, especially at the time of establishing my field experiments, he gave me a lot of guidance and direct support in data collection and dissertation write-up. I cannot forget the support from the entire technical staff of Root Crops Program at NaCRRI during the data collection. Thank you to all the technical staff of Root Crops under the leadership of **Dr. Titus Alicai**. Without the funding from the Bill and Melinda Gates Foundation and UK Aid, I would not have achieved my dream to earn a Ph.D. Thanks to Gates Foundation and UK Aid for funding my Ph.D. I want to thank the entire NextGen Fraternity for the different learning experiences during my Ph.D. Lastly, I want to thank my family and dedicate this work to my late father **Dradri Karlo**, my mother **Amakaru Margret**, and my brother **Jovan Ondia**, who took care of my earlier education expenses. Thank you for the family support to see me reach this level of academic achievement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

GENERAL INTRODUCTION

Center of origin

Cassava (*Manihot esculenta* Crantz), belonging to the family Euphorbiaceae, is known to be a native crop to tropical Amazon regions of Brazil in South America (Olsen and Schaal, 1999). Cassava was introduced to Africa by Portuguese in the 16th century (Cock, 1985). A study by Olsen and Schaal, (1999), reported that the cultivated species of cassava was derived from populations of subspecies, *flabellifolia.* This is based on the similarity of a single-copy nuclear gene glyceraldehyde 3-phosphate dehydrogenase (G3pdh). The amplified fragment length polymorphic (AFLP) marker data of Roa et al. (1997), similarly supported this finding that the cultivated cassava was derived from populations of subspecies, *flabellifolia*.

Reproductive biology of cassava

Cassava or manioc (2*n*=36), considered either a diploid or allopolyploid, is monoecious and largely outcrossing (El-Sharkawy, 2004). Interestingly, the male and female flowers on the same branched panicle mature at different times. The female flowers open 10–14 days before the male ones on the same branch (Halsey et al., 2008), which means flowering and maturation of males and females on the same panicle are not synchronized. From initiation of the first flower buds, the flowering period can continue for more than 2 months, implying that pollen from one flower may fertilize other flowers on the same plant, resulting in self-pollination (Jenning and Iglesias, 2002). Alternatively, the pollen could pollinate flowers on other nearby plants, leading to cross-pollination (Ceballos et al., 2012).

Flowering depends on the plant's genotype and environmental conditions. For example, the early flowering genotypes are known to initiate flower buds within 3 to 5 months after

planting, while the late flowering genotypes take 8 to 10 months. Hence, synchronization of flowering remains a difficult issue in cassava breeding, limiting the number of controlled crosses that can be made (Halsey et al., 2008). Because of the differences in flower initiation and the time required for the seeds to mature, it takes generally no less than a year to obtain seeds of planned crosses (Jennings and Iglesias, 2002). The seeds formed or set are generally few. On average, 1-2 seeds (out of the three possible formed in the trilocular fruit) per pollination are obtained. This limits the amount of the seeds generated from the successful crosses made (Jennings and Iglesias, 2002). In addition, cassava seeds have slow dormancy release, which in turn delays germination. To accelerate germination, the seed dormancy can be broken by exposing seeds to high temperatures (30 – 35 °C) (Ugbede and Hamadina, 2018).

Importance of Cassava

At a global scale, the raising energy demand and climate change has made cassava to be viewed as an alternative source of renewable fuel with greater potential to replace fossil fuel in the developed countries (Kang et al., 2014). The bulk of world trade in cassava is in the form of pellets and chips for animal feed (70%) and the balance mostly in starch and flour for food processing and industrial use (Balagopalan, 2002). Thailand is a dominant supplier of cassava products to world markets, accounting for some 80% of global trade; Vietnam and Indonesia both have a share of about 8%; and a few other countries in Asia, Africa and Latin America provide the least shares in the world market (Balagopalan, 2002).

Cassava is one crop that has the resilience to grow under drought conditions and marginal soils making it a suitable crop for resource poor farming conditions in the face of climate change, especially in sub-Saharan Africa. As such, this crop serves as a major source of carbohydrates for more than 500 million people (El-Sharkawy, 2007). In addition to utilizing

the roots for food, in Africa the boiled cassava leaf is also consumed due its richness in protein, vitamin and other minerals (Montagnac et al., 2009).

In Uganda, cassava ranks as the second most important staple food after "Matooke" (East African highland banana), and mainly produced and consumed in the Eastern and Northern parts of the country (Sserunkuma, 1999). Its roots provide food when consumed fresh (boiled, fried and roasted) or processed into flour for making cassava bread (ASARECA. 2009). In addition to being a food security crop, cassava has gained commercial use in Uganda as a source of raw material for the industrial products such as ethanol, starch, crisps, and animal feed (Otim-Nape et al., 2005).

Production constraints

According to Van Ittersum et al. (2013), yields of staple food crops must be increased substantially over the coming three decades to keep pace with the global food demand driven by the increase in population, especially in sub-Saharan Africa. To achieve this increase in food production, attention must be paid to both abiotic and biotic stresses that impact food crop production. This is compounded by climate change that could affect food supply and increase the risk of hunger globally (Parry et al., 2005). In spite of cassava being a staple food security crop in Sub-Saharan Africa, the yields are still very low (< 12 tons/ha) compared to yield averages of ~ 20 tons/ ha, observed in Asian countries such as Thailand (Nweke, 2004).

In Africa, diverse biotic and abiotic constraints impede cassava production. The economically most important biological constraints are the viral diseases: cassava mosaic disease (CMD) and cassava brown streak disease (CBSD). Both diseases have been known to restrict cassava production since the 1930s (Legg et al., 2014). Two decades ago, less attention was paid to CBSD, because it was endemic to the coastal lowlands of East Africa. Recently,

however, CBSD has attained an epidemic status, covering Eastern, Southern, and Central Africa (Hillocks et al., 2002; Alicai et al., 2007; Legg et al., 2011).

Cassava is also susceptible to bacterial and fungal pathogens, many of which were inadvertently introduced into Africa by earlier scientists. Examples are cassava bacterial blight (CBB), cassava anthracnose (CA) and cassava root rot, which significantly contribute to reduction in cassava production and productivity (Legg et al., 2015). Major pests limiting cassava production include cassava green mites (CGM), *Mononychellus tanajoa*; cassava mealy bug, *Phenacoccus manihoti* Matt.-Ferr; leafhoppers; rodents; and the whitefly, *Bemisia tabaci*, which in addition to being a pest, also acts as the vector for CMD- and CBSD-causing viruses (Maruthi et al., 2005; Omongo et al., 2012). The long cropping cycle of cassava (8 – 24 months) exacerbates many of the disease and pest issues by being hosts for such a long time. Use of clonal propagation via stem cuttings means that infection persists across cropping cycles and is worsened by lack of a formal seed system for ensuring access to clean planting materials (Legg et al., 2014).

Cassava Breeding

(a) Conventional phenotypic recurrent selection

Conventional cassava breeding still dominates most of the breeding programs in Africa. In general, cassava improvement through conventional breeding is a challenging and lengthy process (Rey and Vanderschuren, 2017). Typical conventional breeding schemes involve (i) obtaining F_1 seeds, (ii) seedling evaluation, (iii) first clonal field trial, (iv) preliminary field trials, and (v) advanced field trials. Then two additional years at least are required for adaptability and stability evaluation to support the release of a variety and finally, followed by two to three years for multiplication of planting material (Ceballos et al., 2007; De Oliveira et al., 2012) for distribution. Although conventional recurrent phenotypic selection has provided

the genetic gain that justifies using it in breeding programs (De Oliveira et al., 2012), the time taken to select new parents for recombination and variety release makes it less advantageous than genomics-assisted breeding.

(b) Progress in molecular breeding

Some of the initial steps taken to use molecular breeding tools in cassava included identification of simple sequence repeats (SSR) and random amplified polymorphic DNA (RAPD) markers that were associated with the putative single dominant resistance gene, CMD2 (Raji et al., 2009; Ferguson et al., 2012; Okogbenin et al., 2012). These markers have been used successfully to introgress CMD2-mediated resistance into Latin American germplasm introduced to West Africa cassava to broaden the genetic base for CMD resistance (Okogbenin et al., 2012). The recently developed reference genome sequence for cassava (Prochnik et al., 2012), has further enabled the use of dense single-nucleotide polymorphic (SNPs) markers to finely map the single dominant CMD2 resistant gene in a bi-parental mapping population developed from West African germplasm (Rabbi et al., 2014). Currently, the CMD2 gene that was discovered from a Nigerian landrace (Oliveira Gilmara Alvarenga Fachardo, 2015; Parkes et al., 2015) has been widely used in African cassava breeding programs and in Latin America to breed for CMD resistant clones alongside polygenic recessive CMD1 derived from *Manihot glazovii* at the Amani breeding program in Tanzania (Storey and Nichols, 1938). More recently, Wolfe et al. (2016) identified 13 other genomic regions, including one on chromosome 9 that co-localized with the putative CMD1 locus.

Similarly, quantitative trait loci (QTL) associated with cassava brown streak disease resistance have been identified in Eastern Africa breeding populations through association and bi-parental mapping studies (Kawuki et al., 2016; Masumba et al., 2017; Kayondo et al., 2018).

Another milestone in developing resistance for CBSD has been achieved through cassava genetic transformation with coat proteins of the two CBSD-causing virus species

(Uganda cassava brown streak virus and cassava brown streak virus) in East African germplasm for both improved varieties, as well as the landraces with enhanced yields (Ogwok et al., 2012; Odipio et al., 2014; Beyene et al., 2017; Wagaba et al., 2017). However, the deployment of CBSD resistant transgenic clones continue to face challenges of health and environmental risk perceptions associated with genetically modified crops (Rey and Vanderschuren, 2017).

With the availability of the genomic resources for cassava and low-cost genotyping technologies such as genotyping-by-sequencing (Elshire et al., 2011), cassava breeding is evolving from the traditional phenotypic selection to selecting plants based on their genomic estimated breeding values (GEBVs, genomic selection). Genomic selection (GS), which uses high-density markers to cover the entire genome, was proposed by Meuwissen et al. (2001) as a new method for selection of an individual in a population based on the breeding values. Genomic selection has been reported to offer some advantages over phenotypic selection breeding scheme: (i) genomic selection allows for more cycles of recurrent selection and recombination per unit time than phenotypic selection, (ii) selection is solely based on estimates of marker effects without prior knowledge of the QTL and also captures variation due to loci with small effects (De Oliveira et al., 2012). Another argument put in favor of genomic selection is that genotyping cost will further decrease per sample; on the other hand phenotyping costs do not exhibit the same downward trend, because they are dependent on human resources and agricultural inputs. The cost of these resources have historically been increasing (De Oliveira et al., 2012; Poland and Rife, 2012).

Initial genomic prediction accuracies in cassava were reasonably high for traits that are highly heritable such as dry matter content and cassava mosaic disease (Wolfe et al., 2017), and moderate for cassava brown streak disease (r = 0.24-0.45) (Kayondo et al. 2018). In

contrast, low prediction accuracies were observed for fresh root yield ($r < 0.35$) (Wolfe et al., 2017).

Although the initial results of genomic selection in cassava have been promising, there are still a number of unanswered questions such as the impact of GS on genetic diversity levels, inbreeding rate, influence of G x E on prediction accuracies of traits, as well as predicting a population with a different breeding history. The main objectives of the present study were:

1. To assess genetic variation, inbreeding, and trait correlations across breeding cycles of genomic selection in East African germplasm.
2. To leverage genomic data and environmental covariates for predicting clone performance across environments.
3. To assess genomic prediction accuracies for CBSD resistance in West African germplasm as a pre-emptive breeding strategy.

References

Alicai, T., C.A. Omongo, M.N. Maruthi, R.J. Hillocks, Y. Baguma, R. Kawuki, A. Bua, G.W. Otim-Nape, and J. Colvin. 2007. Re-emergence of Cassava Brown Streak Disease in Uganda. Plant Dis. 91(1): 24–29.

ASARECA, 2009. Generating and scaling out technologies and innovations in cassava for improved livelihood. An annual report.

Balagopalan, C. 2002. Cassava utilization in food, feed and industry. Cassava Biol. Prod. Util. 301–318.

Beyene, G., R.D. Chauhan, M. Ilyas, H. Wagaba, C.M. Fauquet, D. Miano, T. Alicai, and N.J. Taylor. 2017. A Virus-Derived Stacked RNAi Construct Confers Robust Resistance to Cassava Brown Streak Disease. Front. Plant Sci. 7: 1–12.

Ceballos, H., P. Kulakow, and C. Hershey. 2012. Cassava Breeding : Current Status , Bottlenecks and the Potential of Biotechnology Tools. Trop. Plant Biol. 5: 73–87.

Ceballos, H., J.C. Pérez, F. Calle, G. Jaramillo, J.I. Lenis, N. Morante, and J. López. 2007. A new evaluation scheme for cassava breeding at CIAT. : 125–135.

Cock, J.H 1985. Cassava: New potential for a neglected crops. West view Press Bouder, Colorado, USA.

De Oliveira, E.J., M.D.V. de Resende, V. da Silva Santos, C.F. Ferreira, G.A.F. Oliveira, M.S. da Silva, L.A. de Oliveira, and C.I. Aguilar-Vildoso. 2012. Genome-wide selection in cassava. Euphytica 187: 263–276.

El-Sharkawy, M.A. 2004. Cassava biology and physiology. Plant Mol. Biol. 56(4): 481–501.

El-Sharkawy, M.A. 2007. Physiological characteristics of cassava tolerance to prolonged drought in the tropics: Implications for breeding cultivars adapted to seasonally dry and semiarid environments. Brazilian J. Plant Physiol. 19(4): 257–286.

Elshire, R.J., J.C. Glaubitz, Q. Sun, J. a Poland, K. Kawamoto, E.S. Buckler, and S.E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6: 5–13.

Ferguson, M., I. Rabbi, D.J. Kim, M. Gedil, L.A.B. Lopez-Lavalle, and E. Okogbenin. 2012. Molecular Markers and Their Application to Cassava Breeding: Past, Present and Future. Trop. Plant Biol. 5(1): 95–109.

Halsey, M.E., K.M. Olsen, N.J. Taylor, and P. Chavarriaga-Aguirre. 2008. Reproductive biology of cassava (*Manihot esculenta* Crantz) and isolation of experimental field trials. Crop Sci. 48(1): 49–58.

Hillocks, R.J., J.M. Thresh, J. Tomas, M. Botao, R. Macia, and R. Zavier. 2002. Cassava brown streak disease in northern Mozambique. Int. J. Pest Manag. 48: 178–181.

Iglesias, C., and D. J Jennings. 2002. Cassava: biology, production and utilization. Hillocks, R.J., J.M. Thresh (eds), Natural Resources Institute, University of Greenwich, UK, A Bellotti, Centro Internacional de Agricultura Tropical (CIAT), Cali, Colombia, pp 149

Van Ittersum, M.K., K.G. Cassman, P. Grassini, J. Wolf, P. Tittonell, and Z. Hochman. 2013. Yield gap analysis with local to global relevance-A review. F. Crop. Res. 143: 4–17.

Kang, Q., L. Appels, J. Baeyens, R. Dewil, and T. Tan. 2014. Energy-Efficient Production of Cassava-Based Bio-Ethanol. Adv. Biosci. Biotechnol. 5(October): 925–939.

Kawuki, R.S., T. Kaweesi, W. Esuma, A. Pariyo, I.S. Kayondo, A. Ozimati, V. Kyaligonza, A. Abaca, J. Orone, R. Tumuhimbise, E. Nuwamanya, P. Abidrabo, T. Amuge, E. Ogwok, G. Okao, H. Wagaba, G. Adiga, T. Alicai, C. Omongo, A. Bua, M. Ferguson, E. Kanju, and Y. Baguma. 2016. Eleven years of breeding efforts to combat cassava brown streak disease. Breed. Sci. 66: 560–571.

Kayondo, S.I., D.P. Del Carpio, R. Lozano, A. Ozimati, M. Wolfe, Y. Baguma, V. Gracen, S. Offei, M. Ferguson, R. Kawuki, and J.L. Jannink. 2018. Genome-wide association mapping and genomic prediction for CBSD resistance in *Manihot esculenta*. Sci. Rep. 8: 1–11.

Legg, J.P., S.C. Jeremiah, H.M. Obiero, M.N. Maruthi, I. Ndyetabula, G. Okao-Okuja, H. Bouwmeester, S. Bigirimana, W. Tata-Hangy, G. Gashaka, G. Mkamilo, T. Alicai, and P. Lava Kumar. 2011. Comparing the regional epidemiology of the cassava mosaic and cassava brown streak virus pandemics in Africa. Virus Res. 159: 161–170.

Legg, J., E.A. Somado, I. Barker, L. Beach, H. Ceballos, W. Cuellar, J. Lorenzen, J. Lynam, M. Mcmahon, G. Maruthi, D. Miano, K. Mtunda, P. Natwuruhunga, E. Okogbenin, P. Pezo, E. Terry, G. Thiele, M. Thresh, J. Wadsworth, S. Walsh, S. Winter, J. Tohme, and C. Fauquet. 2014. A global alliance declaring war on cassava viruses in Africa. : 231–248.

Legg, J. P., P. L. Kumar, T. Makeshkumar, L. Tripathi, M. Ferguson, E. Kanju, P. Ntawuruhunga, and W. Cuellar. 2015. Cassava Virus Diseases: Biology, Epidemiology, and Management. Advances in Virus Research 91**:**85-142.

Maruthi, M.N., R.J. Hillocks, K. Mtunda, M.D. Raya, M. Muhanna, H. Kiozia, A.R. Rekha, J. Colvin, and J.M. Thresh. 2005. Transmission of Cassava brown streak virus by Bemisia

tabaci (Gennadius). J. Phytopathol. 153: 307–312.

Masumba, E.A., F. Kapinga, G. Mkamilo, K. Salum, H. Kulembeka, S. Rounsley, J. V. Bredeson, J.B. Lyons, D.S. Rokhsar, E. Kanju, M.S. Katari, A.A. Myburg, N.A. van der Merwe, and M.E. Ferguson. 2017. QTL associated with resistance to cassava brown streak and cassava mosaic diseases in a bi-parental cross of two Tanzanian farmer varieties, Namikonga and Albert. Theor. Appl. Genet. 130: 2069–2090.

Meuwissen, T. H. E. , Hayes, B. J., & Goddard, M.E. 2001. Prediction of total genetic value using genome-wide dense markers maps. Genetics 157: 1819–1829.

Montagnac, J. a, C.R. Davis, and S. a Tanumihardjo. 2009. Nutritional Value of Cassava for UseasaStaple Food and Recent Advances for Improvement. Compr. Rev. Food Sci. Food Saf. 8: 181–194.

Nweke. 2004. New challenges in the Cassava tarinsformation in Nigeria and Ghana: Environment and Production Technology Division International Food Policy Research Institute. Food Policy 67: 1–118.

Odipio, J., E. Ogwok, N.J. Taylor, M. Halsey, A. Bua, C.M. Fauquet, and T. Alicai. 2014. RNAi-derived field resistance to Cassava brown streak disease persists across the vegetative cropping cycle. GM Crops Food 5: 16–19.

Ogwok, E., J. Odipio, M. Halsey, E. Gaitán-Solís, A. Bua, N.J. Taylor, C.M. Fauquet, and T. Alicai. 2012. Transgenic RNA interference (RNAi)-derived field resistance to cassava brown streak disease. Mol. Plant Pathol. 13: 1019–1031.

Okogbenin, E., C.N. Egesi, B. Olasanmi, O. Ogundapo, S. Kahya, P. Hurtado, J. Marin, O. Akinbo, C. Mba, H. Gomez, C. De Vicente, S. Baiyeri, M. Uguru, F. Ewa, and M. Fregene. 2012. Molecular marker analysis and validation of resistance to cassava mosaic disease in elite cassava genotypes in Nigeria. Crop Sci. 52(6): 2576–2586.

Oliveira Gilmara Alvarenga Fachardo, de O.E.J. do C.D.C. da S.M.S. 2015. Molecular-assisted selection for resistance to cassava mosaic disease in Manihot. Sci. Agric. 72(6): 520–527.

Olsen, K.M., and B. a. Schaal. 1999. Evidence on the origin of cassava: Phylogeography of *Manihot esculenta*. Proc. Natl. Acad. Sci. 96(May): 5586–5591.

Omongo, C. a., R. Kawuki, A.C. Bellotti, T. Alicai, Y. Baguma, M.N. Maruthi, A. Bua, and J. Colvin. 2012. African Cassava Whitefly, Bemisia tabaci, Resistance in African and South American Cassava Genotypes. J. Integr. Agric. 11: 327–336.

Otim-Nape, G.W., A. Bua, G. Ssemakula, G. Acola, Y. Baguma, S. Ogwal, and R. Van der Grift. 2005. Cassava Development in Uganda a Country Case Study Towards a Global Cassava. A Rev. cassava Africa with Ctry. case Stud. Niger. Ghana, United Repub.

Tanzania, Uganda Benin Proc. Valid. FORUM Glob. CASSAVA Dev. Strateg. Vol. 2: 357.

Parkes, E., M. Fregene, A. Dixon, E. Okogbenin, B. Boakye-Peprah, and M. Labuschagne. 2015. Developing Cassava Mosaic Disease resistant cassava varieties in Ghana using a marker assisted selection approach. Euphytica 203(3): 549–556.

Parry, M., C. Rosenzweig, and M. Livermore. 2005. Climate change, global food supply and risk of hunger. Philos. Trans. R. Soc. B Biol. Sci. 360(1463): 2125–2138.

Poland, J.A., and T.W. Rife. 2012. Genotyping-by-Sequencing for Plant Breeding and Genetics. Plant Genome J. 5(3): 92.

Prochnik, S., P.R. Marri, B. Desany, P.D. Rabinowicz, C. Kodira, M. Mohiuddin, F. Rodriguez, C. Fauquet, J. Tohme, T. Harkins, D.S. Rokhsar, and S. Rounsley. 2012. The Cassava Genome: Current Progress, Future Directions. Trop. Plant Biol. 5: 88–94.

Rabbi, I., M. Hamblin, M. Gedil, P. Kulakow, M. Ferguson, A.S. Ikpan, D. Ly, and J.L. Jannink. 2014. Genetic mapping using genotyping-by-sequencing in the clonally propagated cassava. Crop Sci. 54: 1384–1396.

Raji, A.A., J. V. Anderson, O.A. Kolade, A.G. Dixon, and I.L. Ingelbrecht. 2009. Gene-based microsatellites for cassava *(Manihot esculenta* Crantz): Prevalence, polymorphisms, and cross-taxa utility. BMC Plant Biol. 9: 1–11.

Rey, C., and H.V. Vanderschuren. 2017. Cassava Mosaic and Brown Streak Diseases: Current Perspectives and Beyond. Annu. Rev. Virol. 4(1): annurev-virology-101416-041913.

Roa, A. C., M. Maya, M.C Duque, J.Tohme, A.C., Allem and M.C Bonierbale 1997. AFLP analysis of relationship among cassava and other manihot species. Theoretical and Applied Genetics. 95: 741-750

Sserunkuma. T. 1999. The economic impact of investment in cassava in Uganda. An annual review of cassava research in Uganda.

Storey, H.H., and R.F.W. Nichols. 1938. Studies of the mosaic disease of cassava. Ann. Appl. Biol. 25: 790–806.

Ugbede, E.E., and E.I. Hamadina. 2018. Dormancy in Seeds of Hybrid Cassava Varieties (TMS 98/0505 and TMS 95/0379) Prior to Hardening of Seed Coat. Int. J. Agric. For. 8(2): 98–103.

Wagaba, H., G. Beyene, J. Aleu, J. Odipio, G. Okao-Okuja, R.D. Chauhan, T. Munga, H. Obiero, M.E. Halsey, M. Ilyas, P. Raymond, A. Bua, N.J. Taylor, D. Miano, and T. Alicai.

2017. Field Level RNAi-Mediated Resistance to Cassava Brown Streak Disease across Multiple Cropping Cycles and Diverse East African Agro-Ecological Locations. Front. Plant Sci. 7.

Wolfe, M.D., D.P. Del Carpio, O. Alabi, L.C. Ezenwaka, U.N. Ikeogu, I.S. Kayondo, R. Lozano, U.G. Okeke, A.A. Ozimati, E. Williams, C. Egesi, R.S. Kawuki, P. Kulakow, I.Y. Rabbi, and J.-L. Jannink. 2017. Prospects for genomic gelection in cassava breeding. Plant Genome 10: 1–19.

Wolfe, M.D., I.Y. Rabbi, C. Egesi, M. Hamblin, R. Kawuki, P. Kulakow, R. Lozano, D.P. Del Carpio, P. Ramu, and J.-L. Jannink. 2016. Genome-Wide Association and Prediction Reveals Genetic Architecture of Cassava Mosaic Disease Resistance and Prospects for Rapid Genetic Improvement. Plant Genome 9: 342–356.

CHAPTER 2

GENETIC VARIATION AND TRAIT CORRELATIONS IN EAST AFRICAN CASSAVA BREEDING POPULATION FOR GENOMIC SELECTION

Abstract

Cassava (*Manihot esculenta* Crantz) is a major source of dietary carbohydrates for more than 700 million people living in tropical and sub-tropical parts of the world. However, its long breeding cycle has slowed the rate of genetic gain for target traits. Next generation sequencing has been used to genotype cassava for downstream application of genomic selection with the goal of shortening the cassava breeding cycle. This study aimed at assessing genetic variation, the level of inbreeding and trait correlations in genomic selection breeding cycles. We used phenotypic and genotypic data from the National Crops Resources Research Institute (NaCRRI) foundation population (cycle zero, C_0) and the progeny (cycle one, C_1) derived from crosses of 100 selected C_0 clones as progenitors, both to evaluate and optimize genomic selection. We estimated trait heritability and correlations, population structure, and level of inbreeding in C_0 and C_1. The highest broad-sense heritability (H^2 = 0.95) and narrow-sense heritability (h^2 = 0.81) were recorded for cassava mosaic disease severity and lowest for root weight per plot (H^2 = 0.06 and h^2 = 0.00). We observed the highest genetic correlation (r_g= 0.80) between cassava brown streak disease root incidence, caused by Uganda cassava brown streak and cassava brown streak viruses, measured at seedling and clonal stages of evaluation, suggesting the usefulness of seedling data in predicting clonal performance for cassava brown streak root necrosis. Similarly, high genetic correlations were observed between cassava brown streak disease severity (r_g= 0.83) scored at three and six months after planting (MAP), and cassava mosaic disease caused by African cassava mosaic virus, scored at three and six MAP

(r_g= 0.95), indicating that data obtained on these two diseases at six MAP would suffice. Population differentiation between C_0 and C_1 was not well defined, implying that the 100 selected progenitors of C_1 captured the diversity in the C_0. Overall, the study showed genetic gain for most traits, while maintaining the original genetic diversity in the breeding population as advances were made from C_0 to C_1

This chapter was submitted for publication to Crop Science, Alfred Ozimati, Robert Kawuki, Williams Esuma, Siraj I. Kayondo, Anthony Pariyo, Marnin Wolfe, and Jean-Luc Jannink. 2018. Genetic Variation and Trait Correlations in East African Cassava Breeding Population for Genomic Selection.

Abbreviations

BLUPs, best linear unbiased predictors; GBS, genotyping-by-sequencing; GEBVs, genomic estimated breeding values; GS, genomic selection; C_0, cycle zero; C_1, cycle one; MAF, minor allele frequency; PCA, principal component analysis; SNP, single nucleotide polymorphism; TP, training population; QTL, quantitative trait loci; MAP, months after planting; CBSD, cassava brown streak disease; CBSD3s, cassava brown streak disease severity scored at three months after planting; CBSD3i, cassava brown streak disease incidence at three months after planting; CBSD6s, cassava brown streak disease severity scored at six months after planting, CBSD6i; cassava brown streak disease incidence at six months after planting; CMD, cassava mosaic disease; CBSDRs, cassava brown streak disease root severity assessed at twelve months harvest; CBSDRi, cassava brown streak disease root incidence at twelve months harvest; CMD3s, cassava mosaic disease severity at three months after planting; CMD3i, cassava mosaic disease incidence at three months after planting; CMD6s cassava mosaic disease severity scored at six months after planting; CMD6i, cassava mosaic disease incidence at six months after planting; RTWT, root weight per plot; HI, harvest index; and DMC, dry matter content.

Introduction

Cassava (*Manihot esculenta* Crantz) is a crop that provides staple food for more than 700 million people worldwide (Edgerton, 2009; Burns et al., 2010). Though cassava was domesticated more than 6,000 years ago (Olsen and Schaal, 1999), it has had only a short exposure to formal breeding compared with other staple crops, such as maize (*Zea mays*), rice (*Oryza sativa*) and wheat (*Triticum aestivun L.)* (Iglesias et al., 2004; Fischer and Edmeades, 2010). Formal cassava breeding in Africa only began in the 1930s at the Amani Research Station in Tanzania, where efforts were made to combat epidemics of cassava mosaic disease (CMD) and cassava brown streak disease (CBSD) (Storey and Nichols, 1938). Since then, breeding efforts have yielded substantial genetic improvement in cassava for agronomic traits, including CMD (Jennings and Iglesias, 2002).

However, CBSD has remained a major limitation to cassava production in eastern, central and southern Africa (Alicai et al., 2007; Hillocks and Maruthi, 2015), with lack of resistant varieties amplifying the geographical spread of the disease. The rapid growth of the human population in sub-Saharan Africa and the escalating effects of climate change justify the need for accelerating the rate of genetic gain to increase the productivity of cassava (Burns et al., 2010).

The most commonly used method for breeding cassava remains phenotypic recurrent selection, which requires 8-10 years of evaluation prior to official cultivar release and selection of parents for the next cycle of recombination (Ceballos et al., 2016). The long breeding cycle makes it challenging for the breeders to timely respond to farmers' needs of high yielding and disease resistant cultivars. Fortunately, the availability of relatively cheap next generation sequencing technologies, such as genotyping-by-sequencing (GBS), has made it possible to profile single nucleotide polymorphic (SNP) markers across a genome (Elshire et al., 2011).

This technology enables mapping of quantitative trait loci (QTL) and application of genome-wide predictions, as proposed by Meuwissen et al. (2001).

Genomic selection (GS) involves the prediction of breeding values and selection of parents based on marker-estimated effects, enabling more cycles of selection and recombination per unit time than phenotypic recurrent selection (Bhat et al., 2016). Thus, GS will potentially shorten the breeding cycle of cassava and enable breeders to meet the growing need for improved varieties. However, for GS to be successfully applied in breeding, a number of factors must be considered, including the level of genetic variability and the heritability of the traits for which genome-wide predictions are targeted (Jannink et al., 2010; Muranty et al., 2015).

The National Crops Resources Research Institute (NaCRRI) of Uganda is one of the first cassava breeding programs in Africa to implement GS. Genomic prediction accuracies from the initial training population (C_0) at NaCRRI have been estimated to be reasonably accurate for highly heritable traits, such as dry matter content (DMC) and CMD, but less for low heritability traits, such as fresh root yield (Wolfe et al., 2017). Within the same population (C_0), mean prediction accuracies for CBSD-related traits spanned from 0.24 to 0.43 for CBSD foliar symptoms, and from 0.32 to 0.45 for CBSD root necrosis across a number of tested genomic prediction models (Kayondo et al., 2018). While Wolfe et al. (2017) focused on predicting yield traits and CMD, which are common problems across all cassava breeding programs in Africa, Kayondo et al. (2018) specifically focused on QTL mapping and genomic predictions for CBSD-related traits, a problem facing cassava production only in eastern, central and southern parts of Africa.

In cassava breeding, one of the potential benefits of GS is that selections can be made at the seedling stage, especially for highly heritable traits, for subsequent crossing. Selections at the seedling stage would offer the advantage of reducing the breeding cycle, especially when

the correlation between clonal and seedling performance for the target trait(s) is high. At the International Center for Tropical Agriculture (CIAT), 8-year cassava breeding cycle was reportedly shortened to only three years, when parental selections for subsequent crossing were made at the seedling stage for total carotenoid content (Ceballos et al., 2013). Genetic correlations between seedling and clonal trait expressions have not yet been ascertained in our breeding population; therefore, one of our objectives was to estimate the genetic correlations between seedling and clonal evaluated traits.

Cassava breeding requires selecting for multiple traits to enhance cultivar adoption rate (Barandica et al., 2016). Multi-trait breeding goals are easier to achieve when favorable genetic relationships, arising from linkage or pleiotropy, exist among target traits (Lynch and Walsh, 1998). Phenotypic correlations could be attributed to genetic effects, common environment, or error deviations. On the other hand, an additive genetic correlation between any two traits implies relationship between the breeding values of individuals (Bernardo, 2003). In cassava, undesirable phenotypic and genotypic correlations have been reported for some important traits (Barandica et al., 2016; Esuma et al., 2016; Njoku et al., 2015;). For example, an undesirable negative genetic correlation (r_g = -0.45) between dry matter content and total carotenoid content has been observed in African cassava breeding population (Esuma et al., 2016). In addition to seedling-clonal genetic correlations, it was also of interest to investigate the genetic correlation among clonal evaluated traits.

Furthermore, cassava is known to suffer from inbreeding depression (Rojas et al., 2009; Kaweesi et al., 2014; Ramu et al., 2017). With rapid genomic selection, there is a risk to exacerbate the inbreeding, mainly because of increased selection intensity per unit time. The impact that GS will have on inbreeding in cassava is not yet known, particularly for the NaCRRI GS program. The NaCRRI has recently completed its first cycle (C_1) of GS, including seedling and clonal phenotypic evaluations of a large portion of the C_1 population. This

presents the opportunity to address a number of unanswered questions and assess the progress made so far relative to GS in this population, using both the available C_0 and C_1 datasets. Thus, our overall objective was to estimate genetic parameters to guide routine implementation of GS in East Africa cassava breeding population. Our specific objectives were to: (1) to assess trait variability, genetic diversity and inbreeding level in C_0 and C_1 genotypes that constitute the GS training population at NaCRRI, Uganda; and (2) to examine the phenotypic and genetic correlations for selected agronomic and virus resistance traits evaluated at the seedling and clonal stages.

Materials and Methods

Constitution of C_0 and C_1 populations

In response to the CBSD outbreak in Uganda (Alicai et al., 2007), an initiative was undertaken in 2009 to assemble sources of resistance to facilitate the development of breeding populations for genetic improvement and subsequent on-farm deployment. Accordingly, germplasm was introduced from CIAT, International Institute of Tropical Agriculture (IITA) and Tanzania's national research program. Germplasm from Tanzania was received as botanical seed, whereas materials from CIAT and IITA were introduced as tissue culture plantlets. Hybridizations were made among 52 parents introduced between 2009 and 2010, using a partial diallel mating design. From the progenies generated (full-sibs and half-sibs), 395 clones were selected in 2012 and 2013 to constitute a base population (C_0) for GS. A subset of 100 C_0 clones was selected for hybridization to produce the C_1 population.

In order to select progenitors to generate C_1, we used a selection index. Our selection index included four traits, which collectively represent the major breeding objectives of our program: CBSD root severity (CBSDRs), dry matter content (DMC), harvest index (HI) and root weight per plot (RTWT). As indicated above, our breeding program is implementing genomic selection. We derive genomic-estimated breeding values (GEBVs) for each of the traits mentioned using mixed-model methods described in detail below. Since our selection criteria are already estimates of breeding value, it was not necessary to further account for the difference between phenotypic and genetic variance-covariance (Ceron-Rojas et al., 2015) in constructing our selection index. Instead, we simply mean-centered and variance-standardized the GEBVs for each of the four traits and applied the following formula:

$$\mathrm{SI} = 1 * \mathrm{DMC} + 1 * \mathrm{HI} + 1 * \mathrm{RTWT} - 2 * \mathrm{CBSDRs}$$

The weight of -2 was used for CBSDRs as positive value of the GEBV to indicate worse-than-average disease symptoms.

For genotyping, DNA was extracted from approximately 100 mg of fresh young leaves from each of the C_0 clones. All extractions were done using QIAGEN DNeasy extraction kit and DNA was quantified to ensure the required concentrations for sequencing were obtained. The DNA samples were genotyped using the GBS method described by Elshire et al. (2011). Details of the SNP calling, filtering and imputation pipeline we employed have been provided previously (Hamblin and Rabbi, 2014; Wolfe et al., 2016; Wolfe et al., 2017). Furthermore, the C_0 clones selected to be parents of C_1 were grouped into four clusters, using K-means clustering (Lloyd, 1982), implemented on the realized genomic relationship matrix, which was constructed from GBS SNP markers. During crossing of selected parents, priority was given to between-cluster rather than within-cluster crosses to reduce the risk of inbreeding. Hybridizations and seed handling were conducted using the standard procedure described by Mezzalira et al. (2013).

Field evaluation of C_0 population

In April 2013, a panel of 395 C_0 clones initial exposed to CBSD in Namulonge at seedling and clonal evaluation stages (herein referred to as the training population or TP) was planted at three locations: Namulonge in central Uganda (0° 31' 17.99" N and 32° 36' 32.39" E), Ngetta in northern Uganda (2°14' 50.0" N and 32° 54' 00.0" E), and Kasese in south-western Uganda (0°10' 59.99" N and 30°4' 59.99" E), to assess their agronomic performance and reaction to CMD and CBSD. Importantly, Namulonge is known to be a hotspot for CMD and CBSD (Kaweesi et al., 2014; Pariyo et al., 2015). At each location, single-row plots of 10 plants were established in a 33 x 13 alpha lattice design, with 33 incomplete blocks and two replications. Plant spacing of 1 m x 1 m was adopted within and between rows, whereas blocks were separated by 2 m alleys. No fertilizers were applied during the course of the experiment. Weeding was done manually whenever necessary, and the experiments were entirely rain-fed.

At three and six months after planting (MAP), all plants were assessed for CMD and CBSD shoot symptoms, whereas CBSD root necrosis severity was scored at harvest (12 MAP). Shoot severity for CBSD was assessed on a scale of 1-5 (Hillocks and Thresh, 2000), where 1 = no symptoms; 2 = slight foliar chlorotic leaf mottle with no stem lesions; 3 = foliar chlorotic leaf mottle and blotches with mild stem lesions, but no dieback; 4 = foliar chlorotic leaf mottle and blotches with pronounced stem lesions, but no dieback; and 5 = defoliation with stem lesions and dieback. Foliar incidence for CBSD was computed as a percentage of symptomatic plants per plot. At harvest, all roots in a plot were pooled and assessed individually for CBSD necrosis. Each root was cut transversely into 5-7 pieces, and the cross-sections were scored for necrotic symptoms on a scale of 1-5 (Hillocks and Thresh, 2000), where 1 = no necrosis, 2 = ≤ 5% necrotic; 3 = 6-10% necrotic; 4 = 11-25% necrotic and mild root constriction; and 5 = >25% necrotic and severe root constriction. Root incidence for CBSD was computed as a percentage of necrotic roots per plot. Similarly, CMD severity was scored on a 1-5 scale (IITA, 1990), where 1 = no symptoms; 2 = mild chlorotic pattern across the entire leaf although the leaf appears green and healthy; 3 = moderate mosaic pattern throughout the leaf, narrowing and distortion in the lower one-third of leaflets; 4 = severe mosaic, distortion in two-thirds of the leaflets and general reduction in leaf size; and 5 = severe mosaic distortion in the entire leaf. Foliar incidence for CMD was computed as a percentage of symptomatic plants per plot. On the other hand, plant vigor was evaluated at 3 MAP on an ordinal scale of 3-7, where: 3 = low vigor, 5 = moderate vigor, and 7 = high vigor. At harvest, the aboveground biomass and storage roots for each plot were weighed separately. Harvest index (HI) was computed as a ratio of root weight to total biomass. To measure DMC, 2-5 kg of roots were weighed in air and in water to enable computation of specific gravity, which was subsequently used to estimate DMC, as described by Kawano et al. (1987), as follows:

$$\mathrm{DMC} = 158.3 * (\frac{\mathrm{Wa}}{\mathrm{Wa} - \mathrm{Ww}}) - 142$$

where **Wa** and **Ww** represent weights in air and water, respectively. The ratio in the formula is the specific gravity of the roots. The numbers 158.3 and 142 are the regression coefficient and the intercept, respectively, which were empirically determined by Kawano (1987).

Field evaluation and genotyping of C_1 population

The C_1 seeds generated from crosses among the top 100 progenitors selected from C_0 were processed and germinated under controlled screen house conditions (Mezzalira et al., 2013; Kawano et al., 1980) and the resultant 4,874 C_1 seedlings were transplanted at Namulonge for field evaluation in May 2015. Seedlings were assessed for shoot CMD and CBSD severity at three and six MAP. At 15 MAP (August 2016), plants were harvested and CBSD root necrotic symptoms assessed, as described above. We did not have budget to genotype all C_1 seedlings. For this reason, we decided to cull plants with CMD severity score of 3 or greater as well as those with insufficient stem biomass to generate at least 10 cuttings for subsequent clonal evaluations. We made this decision because CMD is a high heritability trait and easily scored on seedlings. Because we did not cull on the basis of any other trait, the only bias in selection response should be for CMD. Prior to harvesting, leaf samples were collected from 2,113 selected C_1 seedlings for DNA extraction, as described above. Of the 2,113 seedlings, 1,420 were cloned and evaluated at Namulonge. A subset (1,088) of the clones established at Namulonge was evaluated at Serere in eastern Uganda (1° 29' 59.99" N *and* 33° 32' 59.99" E) to capture further variability that might be associated with environmental differences. The clonal trials were established in August 2016, using an augmented design comprising 30-34 plots per block. Each clone was represented by a row of 10 plants and each block contained four checks. Assessment for CMD, CBSD, DMC, and HI was done, as described for C_0 evaluations. Because of missing plots, a total of 1,056 C_1 clones remained for downstream data analyses, of which 432 and 624 were full-sib and half-sib progenies, respectively. Similarly, the top 110 clones were selected from C_1 population as parents to

generate C_2, using the selection index described above and the parents were clustered using SNP markers, as described previously.

Statistical analyses

To enable estimation of genetic variance and further compute broad-sense heritability for traits measured in C_0 and C_1 clonal evaluations, phenotypic data from the two sets of experiments were fitted to a linear mixed model using the *lme4* package for the R statistical computing software (R Development Core Team, 2008). For analysis of C_0 phenotypic data, the following model was fitted:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Z_{clone}c} + \mathbf{Z_{block(rep)}b} + \mathbf{e}$$

where $\boldsymbol{\beta}$ = a vector of fixed effects of locations and grand mean, and $\mathbf{X}$ = the incidence matrix linking observations to those effects; vector $\mathbf{c}$ = a random effect for clones, where $\mathbf{c} \sim N(0, \mathbf{I}\sigma_c^2)$ and $\mathbf{Z_{clone}}$ = the corresponding incidence matrix, and $\mathbf{I}$ = the identity matrix; vector $\mathbf{b}$ = a random effect for blocks nested in replication, such that $\mathbf{b} \sim N(0, \mathbf{I}\sigma_b^2)$ and $\mathbf{Z_{block(rep)}}$ = the corresponding incidence matrix; and $\mathbf{e}$ = residual, such that $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$. Variance components were extracted from the model used to compute broad-sense heritability (H^2) estimates as follows:

$$H^2 = \sigma_c^2/(\sigma_c^2 + \sigma_e^2)$$

where σ_c^2 = clone variance and σ_e^2 = model residual variance. Similarly, we fitted a mixed model for the C_1 trial, including a fixed effect of location and grand mean, a random effect for clones, blocks nested in location and the random residual term. Accordingly, variance components were extracted to compute broad-sense heritability estimates for C_1 clones.

In addition, we fitted a single-step G-BLUP model, first for C_0 and C_1 populations separately, and later combined the two populations for joint analysis. From separate analyses,

we estimated SNP-based heritability (narrow-sense heritability) for all traits, using the formula as described below:

$$h^2 = \sigma_a^2/(\sigma_a^2 + \sigma_e^2)$$

where σ_a^2 = additive genetic variance, h^2 = narrow-sense heritability and σ_e^2 was the model residual variance.

For the combined C_0 and C_1 data, we fitted a mixed model, where the trial location and grand mean were treated as fixed effects, whereas clone effects were considered random, with realized genomic relationship matrix K constructed using *A.mat* function in rrBLUP package for the markers (Endelman, 2011). The GEBVs were extracted for various traits from the G-BLUP model for combined data sets analyzed and averaged for each population (C_0 and C_1). Furthermore, we performed a t-test to compare mean differences in the GEBVs between C_0 and C_1 populations. Boxplots were generated from the GEBVs, using *ggplot* function built in ggplot2 in R, to visualize variability for each trait between the two populations.

Both phenotypic and genetic correlations were estimated for three scenarios: (i) among C_1 traits evaluated at the seedling stage, (ii) between C_1 traits measured at the seedling and at clonal stages, and (iii) combined data for C_0 and C_1 traits evaluated at the clonal stage. To estimate the phenotypic correlations, we used the raw data without accounting for the field trial designs. For estimation of genetic correlations, we first fitted multi-locational models described above for C_0 and C_1 data sets separately. From linear mixed models, best linear unbiased predictors (BLUPs) were extracted for both C_0 and C_1 clones and de-regressed using the formula described by Garrick et al. (2009).

The genomic breeding values were estimated using multivariate G-BLUP with the *emmremlMultivariate* function in the EMMREML package (Akdemir and Okeke, 2015) in R. As only a single observation on each seedling was recorded, a single-step genomic mixed-model was appropriate to fit seedling data for estimation of genetic correlation among C_1 traits

evaluated at the seedling stage in scenario (i). However, to estimate genetic correlations for scenarios (ii) and (iii), we used a two-step procedure. The BLUPs obtained from the first-step analyses described above were de-regressed. The de-regressed BLUPs for C_0 and C_1 were combined into a single dataset and used as response variables in the second-step for fitting the multivariate genomic mixed-models described for seedling data. This approach estimates a genetic variance-covariance matrix for the traits. The trait variance-covariance matrices were converted to genetic correlation matrices with *cov2cor* function in R.

Population structure, inbreeding and genetic diversity

To assess population structure, genetic diversity and inbreeding among C_0 and C_1, we used 46,760 SNP markers in both C_0 and C_1 populations. The markers were filtered to have a minor allele frequency (MAF) ≥ 0.01 and formatted as a dosage matrix, with SNP genotypes coded as -1, 0 or +1. The realized genomic relationship matrix (K) was constructed with this dosage matrix as input using the *A.mat* function in the rrBLUP package (Endelman, 2011). Principal component analysis (PCA) was conducted on K using the *prcomp* function in R. The first two principal components (PCs) were used to visualize population structure. The mean of the diagonals of the matrix K is known to be proportional to the inbreeding coefficient (1+F) (Endelman and Jannink, 2012). Therefore, we used the average of the diagonal elements of K as a proxy to measure inbreeding coefficient. On the other hand, we used the average of the off-diagonal elements of K as a measure of genetic diversity. These averages were computed separately for the C_0 and C_1 populations.

Results

Heritability estimates and mean GEBVs of C_1 and C_0 clones

Estimates of broad-sense heritability (H^2) for foliar CBSD scored at three MAP and six MAP (CBSD3s, CBSD3i, CBSD6s, and CBSD6i) ranged from 0.28 for CBSD3s to 0.47 for CBSD3i, in C_0 base population, whereas the estimates of the broad-sense heritability varied from 0.44 for CBSD6s to 0.59 for CBSD6s in C_1 base population (Table 2. 1). In general, broad-sense heritability for CBSD root necrosis was higher for C_1 (0.45 for CBSDRs and 0.50 for CBSDRi) than for C_0 (0.38 for CBSDRs and 0.37 for CBSDRi) base population. On the other hand, estimates of broad-sense heritability for foliar CBSD ranged from 0.26 for CBSD3s to 0.49 for CBSD3i among selected parents out of C_0, while the broad-sense heritability ranged from 0.52 for CBSD6i to 0.69 for CBSD3i among the selected parents out of C_1. Overall, the broad-sense heritability estimates of CBSD root necrosis were higher for C_1 (0.70 for CBSDRs and 0.63 for CBSDRi) selected as parents than for C_0 (0.29 for CBSDRs and 0.39 for CBSDRi) clones selected as progenitors to generate the C_1 population.

Broad-sense heritability for cassava mosaic disease, an important plant health trait, varied from 0.50 for CMD6i to 0.60 for CMD3s in C_0 base population, whereas the broad-sense heritabilities in C_0 base population ranged from 0.77 for CMD6i to 0.81 for CMD6s for C_1 (Table 2. 1). Meanwhile, broad-sense heritability ranged from 0.47 for CMD6i to 0.61 for CMD6s, and 0.08 for CMD6i to 0.95 for CMD3s for selected parents from C_0 and C_1 base populations.

Broad-sense heritability estimates for HI ranged from 0.20 to 0.40 for C_0 and C_1 populations, and their selected progenitors (Table 2. 2). The broad-sense heritability estimates were generally low for root DMC (≤ 0.18) and root weight per plot (≤ 0.24), among C_0, C_1, and selected parents out of C_0. In contrast, moderate broad-sense heritability estimates of 0.49

and 0.30 were observed, respectively, for DMC and root weight per plot for selected C_1 clones as parents (Table 2. 3).

Estimates of narrow-sense heritability (h^2), also referred to as "SNP-based heritability," for foliar CBSD ranged from 0.27 for CBSD3s to 0.53 for CBSD3i among the C_0 base population, while estimates of narrow-sense varied from 0.46 for CBSD6s to 0.59 for CBSD3i for parents selected from C_1 base population (Table 2. 4). For CBSD root necrosis, in general, we observed higher narrow-sense heritability estimates for C_0 (0.43 for CBSDRs and 0.44 for CBSDRi) than for C_1 (0.06 for CBSDRs and 0.13 for CBSDRi) population. On the other hand, narrow-sense heritability for foliar CBSD ranged from 0.47 for CBSD3s to 0.72 for CBSD3i in C_0, and from 0.57 for CBSD3s to 0.68 for CBSD6i in C_1 for selected progenitors (Table 2. 5). Similar to the base populations, estimates of narrow-sense heritability were higher for CBSD root necrosis in C_0 (0.54 for CBSDRs and 0.65 for CBSDRi) than those in C_1 (0.21 for CBSDRs and 0.31 for CBSDRi) for selected parents.

For CMD, we recorded relatively high narrow-sense heritability estimates, ranging from 0.62 for CMD6s to 0.78 for CMD3i for C_0 base population and from 0.44 for CMD6i to 0.59 for CMD3s for C_1 base population. Meanwhile, narrow-sense heritability varied between 0.72 for CMD6s and 0.82 for CMD3i in C_0, and between 0.05 for CMD6i and 0.23 for CMD3i in C_1 for selected parents from C_1. We observed higher SNP-based heritability estimates for HI in C_0 base population ($h^2 = 0.48$) and their selected progenitors ($h^2 = 0.67$) than those for C_1 base population ($h^2 = 0.18$) and their selected parents ($h^2 = 0.11$). Generally, low ($h^2 \leq 0.36$) SNP-based heritability estimates were recorded for DMC and RTWT for both C_0 and C_1 populations, except for C_1 clones selected as parents, where SNP-based heritability estimate was relatively high ($h^2 = 0.79$) for DMC (Table 2. 6).

We compared the average breeding values of C_0 and C_1 clones (Table 2. 1). For CBSD in general, the C_1 clones had better average breeding values (lower disease) than the C_0 clones.

Further, t-test of mean differences between C_0 and C_1 GEBVs for CBSD6s and CBSDRs, revealed highly significant differences ($P \leq 0.001$) between the two populations. For CMD, C_0 clones exhibited better performance (lower disease) than C_1; however, the mean differences between the GEBVs for the two populations were non-significant for both CMD3s and CMD6s (Table 2. 7). Similar trend of non-significant average difference in GEBVs was observed for RTWT between C_0 and C_1 clones. For dry matter content, the C_1 clones had significantly ($P \leq 0.001$) higher average GEBVs than C_0 clones (Table 2. 8).

Table 2. 9: Heritability estimates and mean genomic estimated breeding values (GEBVs) for traits measured at clonal evaluation stage

	C_0 base Population†		Selected parents out of C_0‡		C_1 base Population§		Selected parents out of C_1¶		C_0 base Population	C_1 base Population	C_0 vs C_1 base Populations
Traits#	**H^2††**	**h^2‡‡**	**H^2**	**h^2**	**H^2**	**h^2**	**H^2**	**h^2**	**GEBVs§§**	**GEBVs**	**t-test C_0 vs C_1**
CBSD3s	0.28	0.27	0.26	0.47	0.55	0.57	0.57	0.57	0.04	-0.07	0.11^{ns}
CBSD3i	0.47	0.53	0.49	0.72	0.56	0.59	0.69	0.60	1.39	-0.68	2.04^{ns}
CBSD6s	0.32	0.32	0.41	0.59	0.44	0.47	0.65	0.65	0.05	-0.02	0.07^{***}
CBSD6i	0.35	0.36	0.34	0.49	0.59	0.46	0.52	0.68	3.11	-1.53	4.64^{***}
CBSDRs	0.38	0.43	0.29	0.54	0.45	0.06	0.70	0.21	0.09	-0.04	0.13^{***}
CBSDRi	0.37	0.44	0.39	0.65	0.50	0.13	0.63	0.31	2.95	-1.44	4.39^{***}
CMD3s	0.51	0.70	0.49	0.81	0.81	0.59	0.95	0.20	-0.03	0.01	0.04^{ns}
CMD3i	0.60	0.78	0.50	0.81	0.78	0.50	0.59	0.23	-2.16	1.05	3.21^{**}
CMD6s	0.56	0.62	0.61	0.72	0.81	0.55	0.25	0.21	-0.02	0.01	0.03^{ns}
CMD6i	0.50	0.65	0.47	0.77	0.77	0.44	0.08	0.05	-1.53	0.74	2.27^{*}
HI	0.36	0.48	0.40	0.67	0.36	0.18	0.20	0.11	0.01	-0.01	0.02^{**}
RTWT	0.24	0.04	0.06	0.17	0.14	0.00	0.30	0.30	0.02	-0.01	0.03^{ns}
DMC	0.11	0.12	0.07	0.06	0.18	0.08	0.49	0.79	-0.22	0.11	0.33^{***}

*, **, ***Significant GEBVs between C_0 and C_1 base populations at 0.05, 0.01, and 0.001 probability level, respectively; †Initial set of 395 clones for training genomic prediction model (C_0 base Population); ‡Progenitors selected (100 clones) from initial training population to generate genomic selection cycle one population.; §Second set of 1056 clones referred to as genomic selection cycle one population (C_1 base Population); ¶Progenitors selected (110 clones) from genomic selection cycle one population to generate cycle two; #CBSD3s, cassava brown streak disease severity assessed at three months after planting; CBSD3i, cassava brown streak disease incidence at three months after planting; CBSD6s, cassava brown streak disease severity assessed at six months after planting; CBSD6i, cassava brown streak disease incidence at six months after planting; CBSDRs, cassava brown streak disease root severity at 12 months after planting; CBSDRi, cassava brown streak disease root incidence at 12 months after planting; CMD3s, cassava mosaic disease severity assessed at three months after planting; CMD3i, cassava mosaic disease incidence at three months after planting; CMD6s, cassava mosaic disease severity assessed at six months after planting; CMD6i, cassava mosaic disease incidence at six months after planting; HI, harvest index; RTWT, root weight per plot; and DMC, dry matter content; †Broad-sense heritability estimates for C_0 and C_1 base populations and their selected progenitors; ‡‡SNP-based heritability estimates (narrow-sense heritability) for C_0 and C_1 base populations and their selected progenitors; §§Mean genomic estimated breeding values (GEBVs).

Furthermore, using boxplots to compare the variation in GEBVs between C_1 and C_0 populations (Figure 2. 1), a general trend of lower CBSD incidences and severities in C_1 than in C_0 was observed for disease assessments at three, six and 12 MAP, with much reduced variability for C_1 clones. For CMD, HI, and DMC, the level of variation for the GEBVs was relatively similar for C_0 and C_1 clones, whereas C_0 had more variability in their GEBVs than C_1 for RTWT and plant vigor (Figure 2. 2).

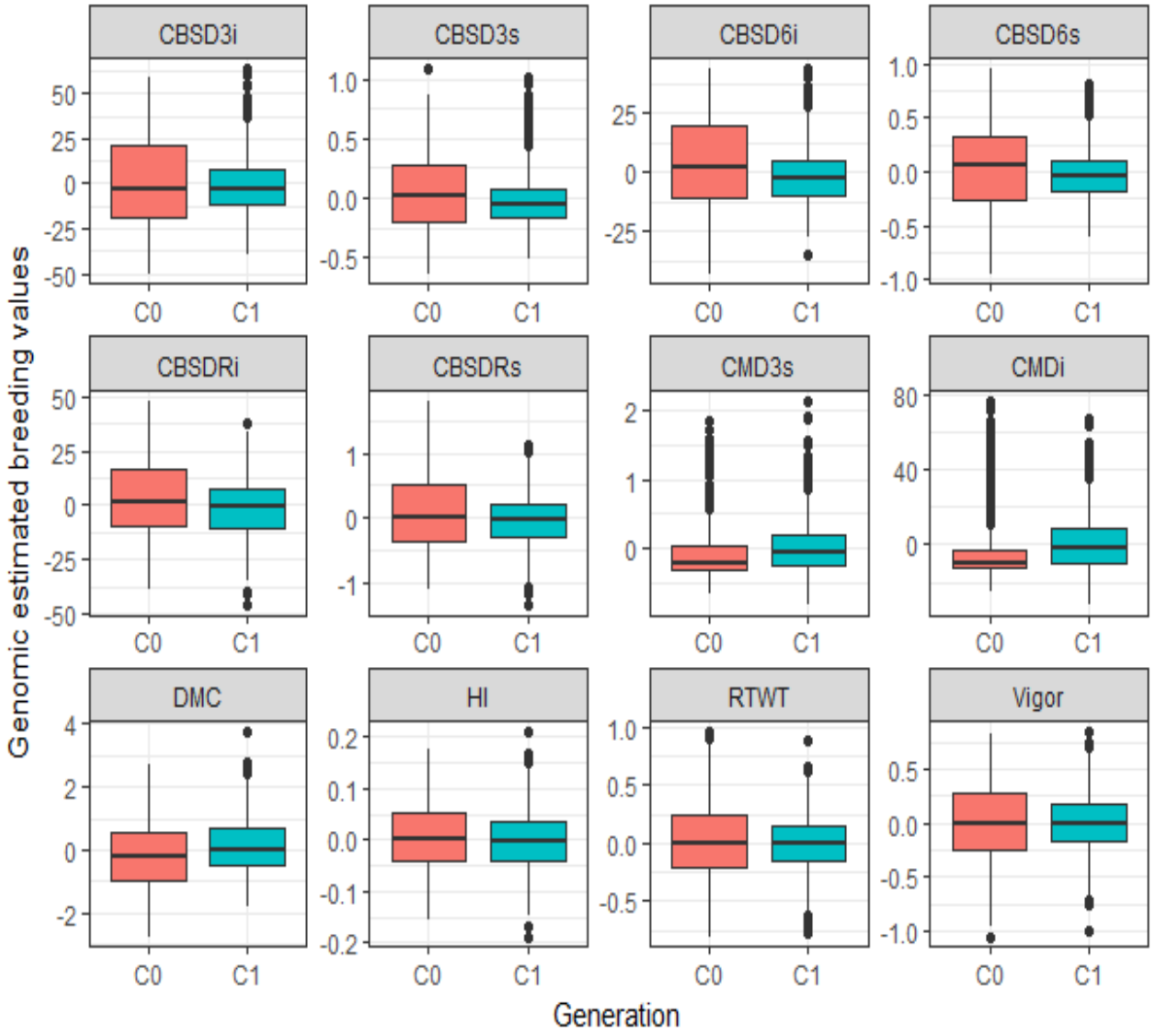


Figure 2. 3: Boxplots showing variability in genomic estimated breeding values for selected plant health and agronomic traits of cycle zero (C_0) and cycle one (C_1) populations evaluated at clonal stage

Genetic and phenotypic correlations among C_0 and C_1 clones

Because root weight for cassava can be estimated reasonably only from the clonal-stage evaluations, the seedling evaluation focused on correlations among plant health traits. Highly positive phenotypic and genetic correlations (r_p= 0.88 and r_g = 0.94, respectively) were recorded between seedling CBSDRs and CBSDRi (Table 2. 10). In general, there were low pair-wise genetic correlations observed among C_1 seedling traits (plant vigor, CBSD6s and CMD6s), varying from r_g = -0.14 to r_g = 0.24 (Table 2. 11).

Table 2. 12: Phenotypic (lower diagonal) and genetic (upper diagonal) correlations among C1 seedling traits

Traits#	Vigor-ST	CBSD6s-ST	CBSDRs-ST	CBSDRi-ST	CMD6s-ST
Vigor-ST	-	0.04	0.12	0.11	-0.02
CBSD6s-ST	-0.04	-	0.24	0.27	0.10
CBSDRs-ST	0.05	0.03	-	0.94	-0.01
CBSDRi-ST	0.05	0.03	0.88	-	-0.14
CMD6s-ST	-0.04	0.12	-0.03	-0.05	-

All the phenotypic and genetic correlations (r_p and r_g) $\geq$ 0.1 in absolute values were significant ($P \leq 0.05$) at an individual test level; †Vigor-ST, seedling plant vigor; CBSD6s-ST, seedling cassava brown streak disease severity assessed at six months after planting (MAP); CBSDRs-ST, seedling cassava brown streak disease root severity assessed at 12 MAP; CBSDRi-ST, seedling cassava brown streak disease root incidence assessed at 12 MAP; and CMD6s-ST, and seedling cassava mosaic disease severity assessed at six MAP

Results for phenotypic correlations between seedling and clonal evaluations are presented in Table 2. 13 and 2. 14. We recorded moderate to high, positive phenotypic and genetic correlations between CBSDRs scored at seedling stage and other CBSD related-traits, assessed at clonal stage, notable of which included: a) CBSD3s (r_p = 0.34 and r_g = 0.23) and CBSD3i (r_p = 0.35 and r_g = 0.30); and b) CBSDRs (r_p = 0.39 and r_g = 0.70), and CBSDRi (r_p = 0.36 and r_g = 0.77). We observed a similar trend for the phenotypic and genetic correlations between CBSDRi scored at seedling and other CBSD- related traits, with the highest correlation (r_g= 0.80) observed between CBSDRi measured at seedling and CBSDRi at clonal stage (Table 2. 15 and 2. 16). Unexpectedly, CMD6s with high heritability estimates had low phenotypic correlation (r_p =0.08) observed between seedling and clonal stages. On the other

hand, a negative genetic correlation (r_g = -0.46) was observed between seedling plant vigor and harvest index measured at clonal stage.

Table 2. 17: Phenotypic correlations for traits measured at seedling and at clonal evaluation stages

Traits†	Vigor-ST	CBSD6s-ST	CBSDRs-ST	CBSDRi-ST	CMD6s-ST
CBSD3s-CT	-0.12	-0.08	0.34	0.29	0.01
CBSD3i-CT	-0.13	-0.08	0.35	0.30	0.02
CBSD6s-CT	-0.08	-0.07	0.26	0.25	0.06
CBSD6i-CT	-0.08	-0.08	0.25	0.24	0.08
CBSDRs-CT	-0.08	0.05	0.39	0.35	-0.11
CBSDRi-CT	-0.09	0.04	0.36	0.36	-0.05
CMD3s-CT	0.03	-0.06	0.01	-0.02	0.05
CMD3i-CT	0.04	-0.07	0.02	-0.01	0.04
CMD6s-CT	0.05	-0.06	0.02	-0.01	0.08
CMD6i-CT	0.07	-0.07	0.01	-0.02	0.11
HI-CT	-0.01	0.01	-0.18	-0.12	0.00
RTWT-CT	0.08	0.03	-0.07	-0.05	0.00
DMC-CT	0.07	-0.04	-0.11	-0.07	0.01
Vigor-CT	0.12	0.00	-0.08	-0.07	0.00

All the phenotypic correlations (r_p) $\geq$ 0.24 in absolute values were significant ($P \leq 0.05$) at an individual test level.
†Vigor-ST, seedling plant vigor; CBSD6s-ST, seedling cassava brown streak disease severity assessed at six months after planting; CBSDRs-ST, seedling cassava brown streak disease root severity at 12 months after planting; CBSDRi-ST, seedling cassava brown streak disease root incidence at 12 months after planting; CMD6s-ST, seedling cassava mosaic disease severity assessed at six months after planting; CBSD3s-CT, clonal cassava brown streak disease severity assessed at three months after planting; CBSD3i-CT,clonal cassava brown streak disease incidence at three after planting; CBSD6s-CT, clonal cassava brown streak disease severity assessed at six months after planting; CBSD6i-CT, clonal cassava brown streak disease incidence at six months after planting; CBSDRs-CT, clonal cassava brown streak disease root severity at 12 months after planting; CBSDRi-CT, clonal cassava brown streak disease root incidence at 12 months after planting; CMD3s-CT, clonal cassava mosaic disease severity scored at three months after planting; CMD3i-CT, clonal cassava mosaic disease incidence at three months after planting; CMD6s-CT, clonal cassava mosaic disease severity scored at six months after planting; CMD6i-CT, clonal cassava mosaic disease incidence at six months after planting; HI-CT, clonal harvest index; RTWT-CT, clonal root weight per plot; DMC-CT, clonal dry matter content; and Vigor-CT, clonal plant vigor

Table 2. 18: Genetic correlations for traits measured at seedling and at clonal evaluation stage

Traits†	Vigor-ST	CBSD6s-ST	CBSDRs-ST	CBSDRi-ST
Vigor-CT	0.04	0.01	-0.02	-0.08
CBSD3s-CT	0.05	-0.25	0.23	0.19
CBSD3i-CT	0.08	-0.27	0.30	0.25
CBSD6s-CT	-0.05	-0.14	0.34	0.29
CBSD6i-CT	0.00	-0.19	0.34	0.29
CBSDRs-CT	-0.10	0.31	0.70	0.73
CBSDRi-CT	0.07	0.31	0.77	0.80
CMD3s-CT	0.21	-0.23	0.02	-0.02
CMD3i-CT	0.31	-0.18	-0.05	-0.09
CMD6s-CT	0.42	-0.30	0.05	0.01
CMD6i-CT	0.59	-0.25	-0.02	-0.06
HI-CT	-0.46	0.02	-0.07	-0.12
RTWT-CT	0.04	0.17	-0.11	0.02
DMC-CT	-0.01	0.06	-0.30	-0.31

All the genetic correlations (r_g) $\geq$ 0.27 in absolute values were significant ($P \leq 0.05$) at an individual test level; †Vigor-ST, seedling plant vigor; CBSD6s-ST, seedling cassava brown streak disease severity assessed at six months after planting; CBSDRs-ST, seedling cassava brown streak disease root severity at 12 months after planting; CBSDRi-ST, seedling cassava brown streak disease root incidence at 12 months after planting; CBSD3s-CT, clonal cassava brown streak disease severity assessed at three months after planting; CBSD3i-CT,clonal cassava brown streak disease incidence at three after planting; CBSD6s-CT, clonal cassava brown streak disease severity assessed at six months after planting; CBSD6i-CT, clonal cassava brown streak disease incidence at six months after planting; CBSDRs-CT, clonal cassava brown streak disease root severity at 12 months after planting; CBSDRi-CT, clonal cassava brown streak disease root incidence at 12 months after planting; CMD3s-CT, clonal cassava mosaic disease severity scored at three months after planting; CMD3i-CT, clonal cassava mosaic disease incidence at three months after planting; CMD6s-CT, clonal cassava mosaic disease severity scored at six months after planting; CMD6i-CT, clonal cassava mosaic disease incidence at six months after planting; HI-CT, clonal harvest index; RTWT-CT, clonal root weight per plot; DMC-CT, clonal dry matter content; and clonal plant vigor.

Finally, we examined phenotypic and genetic correlations among C_0 and C_1 traits evaluated at clonal stage (Table 2. 19). We recorded the high phenotypic and genetic correlations ranging from 0.79 to 0.98 between disease severity and incidence scored within the same time point, i.e., at 3, 6, and 12 MAP for both CMD and CBSD. Similarly, we observed high phenotypic and genetic correlations, ranging from 0.51 to 0.95 between foliar disease severities scored at 3 and 6 MAP for both CMD and CBSD. However, there were notably low phenotypic and genetic correlations between CBSD3s and CBSDRs (r_p = 0.13 and r_g = -0.12), CBSD6s and CBSDRs (r_p = 0.23 and r_g = -0.03). Furthermore, we consistently observed

negative genetic correlations ranging from -0.32 to -0.17 between disease traits and root weight per plot, similar to genetic correlations between DMC and CBSDRs (r_g = -0.64). However, HI had a positive phenotypic and genetic correlations (r_g = 0.4 and r_g = 0.54) with root weight per plant. (Table 2. 20). Importantly, all the phenotypic and genetic correlations (r_g) $\geq$ 0.2 in absolute values were significant ($P \leq 0.05$) at an individual test level.

Table 2. 21: Phenotypic (lower diagonal) and genetic (upper diagonal) correlations among C_0 and C_1 clonal evaluated traits

Traits#	Vigor	CBSD3s	CBSD3i	CBSD6s	CBSD6i	CBSDRs	CBSDRi	CMD3s	CMD3i	CMD6s	CMD6i	HI	RTWT	DMC
Vigor	-	-0.03	-0.06	-0.05	-0.13	0.03	0.04	-0.22	-0.25	-0.21	0.21	-0.02	0.35	0.05
CBSD3s	-0.01	-	0.95	0.83	0.84	-0.12	0.01	-0.04	-0.07	-0.06	-0.06	0.15	-0.20	-0.11
CBSD3i	-0.01	0.79	-	0.90	0.90	-0.15	0.00	-0.08	-0.11	-0.12	-0.10	0.21	-0.17	-0.07
CBSD6s	0.01	0.51	0.52	-	0.95	-0.03	-0.03	-0.28	-0.13	-0.30	-0.29	0.32	-0.21	-0.09
CBSD6i	-0.02	0.49	0.53	0.82	-	0.11	0.11	-0.13	-0.12	-0.12	-0.12	0.26	-0.27	-0.11
CBSDRs	0.01	0.13	0.11	0.23	0.17	-	0.95	-0.22	-0.19	-0.15	-0.20	-0.37	-0.20	-0.64
CBSDRi	0.01	0.15	0.13	0.24	0.18	0.92	-	-0.31	-0.28	-0.25	-0.28	-0.26	-0.18	-0.58
CMD3s	-0.11	-0.06	-0.11	-0.12	-0.09	-0.10	-0.10	-	0.98	0.95	0.95	-0.46	-0.29	0.01
CMD3i	-0.10	-0.09	-0.12	-0.14	-0.11	-0.10	-0.10	0.91	-	0.94	0.95	0.44	-0.32	0.00
CMD6s	-0.13	0.02	-0.05	-0.03	0.00	-0.05	-0.06	0.70	0.68	-	0.97	-0.43	-0.31	-0.05
CMD6i	-0.13	-0.01	-0.05	-0.05	-0.02	-0.07	-0.06	0.70	0.71	0.90	-	-0.41	-0.24	-0.05
HI	0.17	0.00	0.02	0.01	0.03	-0.24	-0.22	-0.09	-0.09	-0.13	-0.12	-	0.41	0.23
RTWT	0.21	-0.03	0.01	0.00	0.01	-0.22	-0.22	-0.09	-0.08	-0.09	-0.08	0.54	-	0.19
DMC	0.22	-0.17	-0.15	-0.28	-0.24	-0.39	-0.36	0.04	0.09	-0.07	-0.04	0.42	0.16	-

The phenotypic and genetic correlations (r_p & r_g) $\geq$ 0.2 in absolute values were significant ($P \leq 0.05$) at an individual test level;†Vigor, plant vigor scored at three months after planting; CBSD3s, cassava brown streak disease severity scored at three months after planting; CBSD3i, cassava brown streak disease incidence at three months after planting; CBSD6s, cassava brown streak disease severity scored at six months after planting; CBSD6i, cassava brown streak disease incidence at six months after planting; CBSDRs, cassava brown streak disease root severity at 12 months after planting; CBSDRi, cassava brown streak disease root incidence at 12 after planting; CMD3s, cassava mosaic disease scored at three months after planting; CMD3i, cassava mosaic disease incidence at three months after planting; CMD6s, cassava mosaic disease severity scored at six months after planting; CMD6i, cassava mosaic disease incidence at six months after planting; HI, harvest index; RTWT, root weight per plot; and DMC, dry matter content.

Population structure and level of inbreeding in C_0 and C_1 clones

Based on PCA, there was no clear genetic differentiation between C_0 and C_1 populations. Indeed, majority of the total genetic variation (49%) in C_0 and C_1 populations was explained by the first PC, with 13% attributed to PC2 (Figure 2. 4). Further plots of the loadings (eigenvector coefficients) for each marker on PC1 and PC2 against marker position along the 18 cassava chromosomes, revealed that markers affecting PC1 and PC2 most strongly were on the first and fourth chromosome, respectively (Figure 2. 5a and 2. 6b).

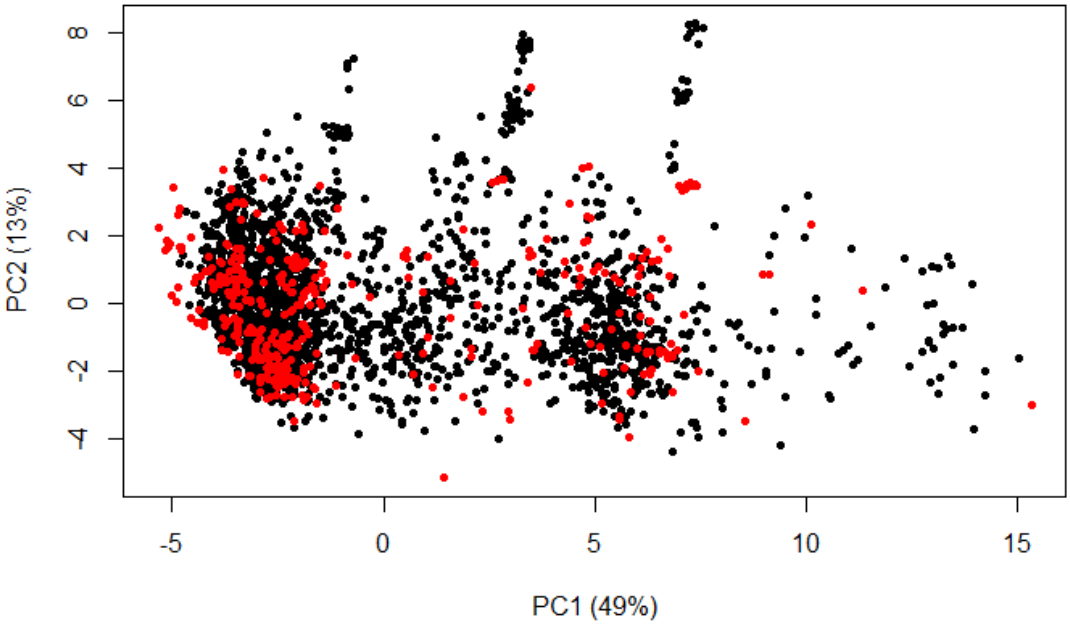


Figure 2. 7: Population structure from a plot of Eigen values of PC1 against PC2, using realized genomic relationship matrix for C_0 and C_1 populations. The C_0 population (red) comprised 395 individuals and C_1 (black) comprised 1056 clones. The population structure was estimated from kinship matrix constructed, using 46,760 SNP markers, filtered at minor allele frequency (MAF) $\geq$ 0.01.

.

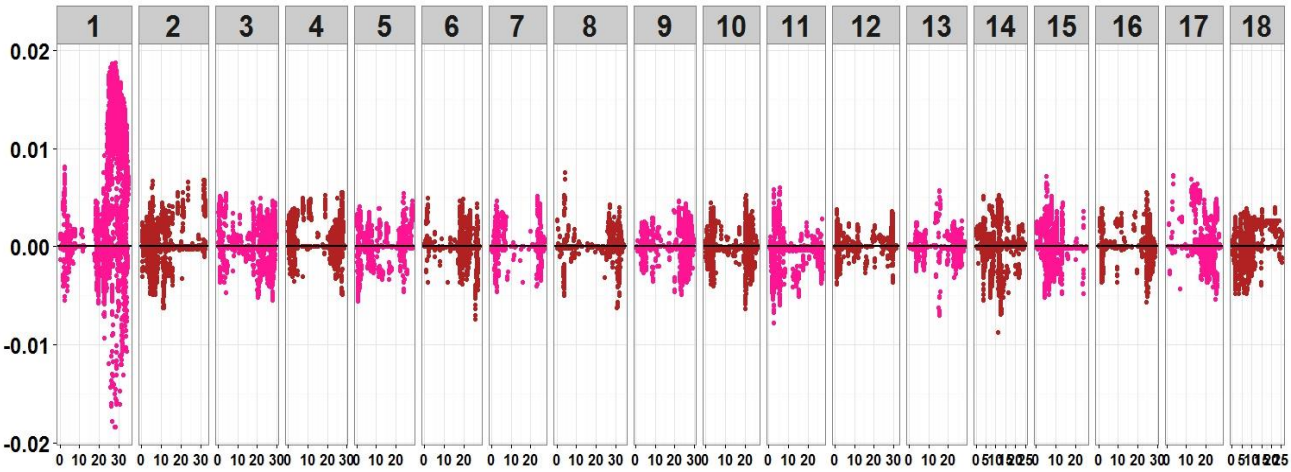


Figure 2. 8a: A plot of the loadings (Eigen vector coefficients) for each marker on PC1 against marker position along the 18 cassava chromosomes. Markers affecting PC1 most strongly loaded on the first chromosome

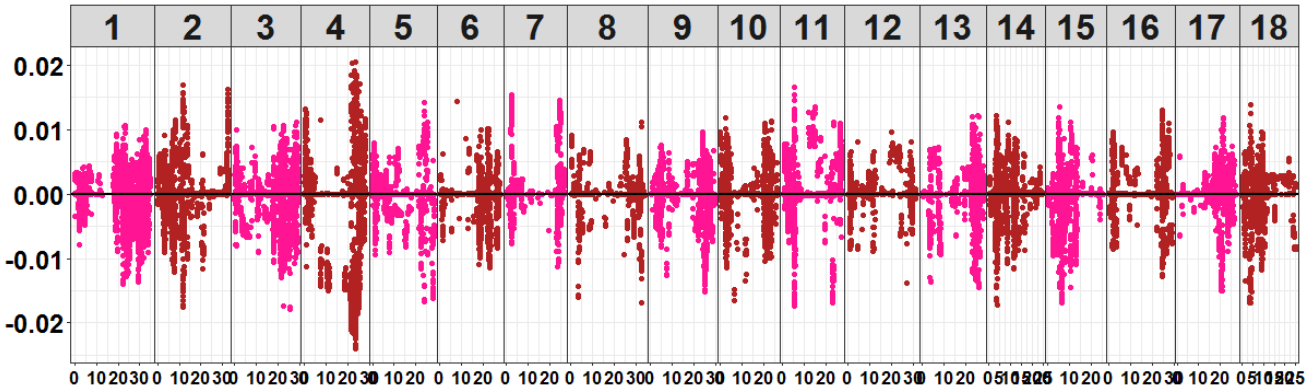


Figure 2. 3b: A plot of the loadings (Eigen vector coefficients) for each marker on PC2 against marker position along the 18 cassava chromosomes**.** Markers explaining the largest variation for PC2 loaded on chromosome four.

The means of the diagonals of the kinship matrix, which is proportional to one plus the inbreeding coefficient (1+F) were 0.904 in C_0 and 0.708 for C_1 (Figure 2. 9a). Density plots of the off-diagonal elements of the kinship matrix indicated that the degree of variability in relatedness was similar in the C_0 and C_1 (Figure 2. 4b).

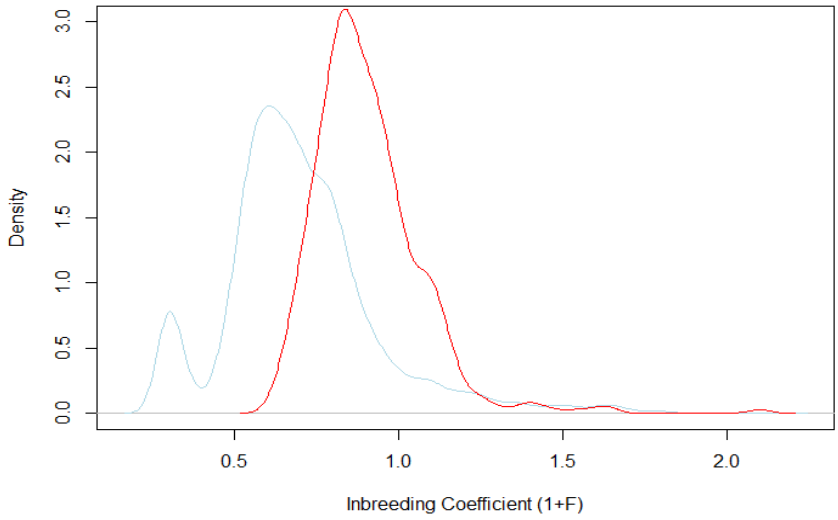


Figure 2. 10a: Density plots generated from the diagonal elements of realized genomic relationship matrix, as measures of inbreeding levels for C_0 and C_1 populations. The density plots indicated that there was less inbreeding in C_1 (Light blue) than in C_0 (Red) clones

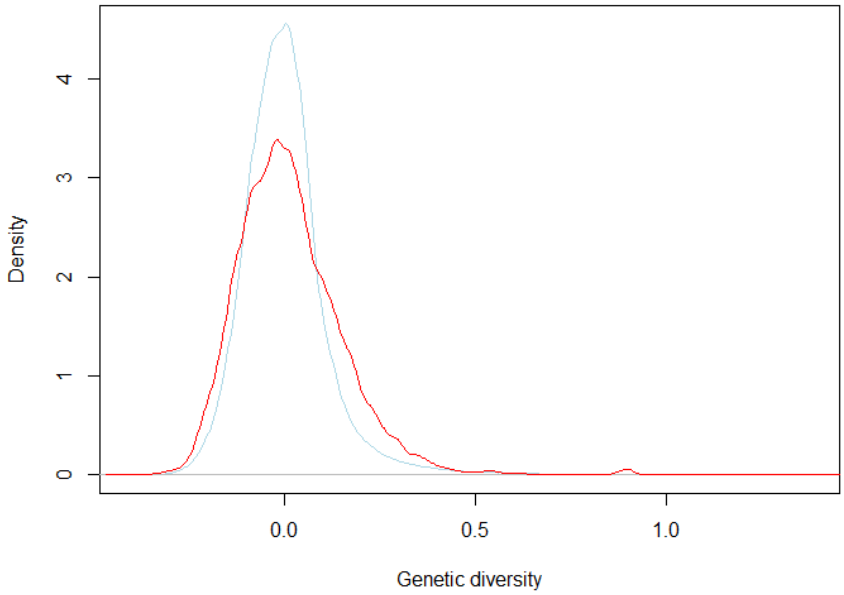


Figure. 2 4b: Density plots, generated from the off-diagonal elements of realized genomic relationship matrix, as a measure of genetic diversity in C_0 (red) and C_1 (light blue) clones for C_0 and C_1 populations. Both C_0 and C_1 had similarity in diversity estimated from 46,760 SNP markers, implying the original diversity in C_0 base population was captured by the selected parents that generated C_1 population.

Discussion

In this paper, we highlight progress that has been made towards increasing the productivity of cassava for the benefit of communities that depend on it. Notable production obstacles being addressed include susceptibility to CMD and CBSD, low yields of both fresh root and dry matter. In response to these challenges, the cassava breeding program based at NaCRRI initiated efforts to implement genomic selection to accelerate the breeding cycle of cassava (Wolfe et al., 2016, 2017). This paper, therefore, aimed at examining genetic variability and correlations in genomic selection populations (C_0 and C_1) as well as their respective selected progenitors. Our primary traits of focus included diseases (CMD and CBSD) and yield components (DMC, HI, and RTWT).

Heritability estimates and mean GEBVs of C_1 and C_0 clones

The relatively high heritability for CMD has already facilitated identification of resistant cultivars from various breeding programs through phenotypic selection (Thresh and Cooter, 2005; Egesi et al., 2007; Kawuki et al., 2016). In general, phenotypic selection would be cost effective to select for highly heritable traits, such as CMD. However, to produce desirable cultivars, our breeding program seeks to improve a number of traits, alongside CMD, many of which are quantitative traits, such as yield and CBSD resistance with low to moderate heritability. It is mainly for these yield and CBSD resistance that we employ GS, using a selection index to simultaneously improve all focal traits.

Moderate to high estimates of broad-sense heritability for foliar and root severities were registered for CBSD (ranging from 0.26 to 0.70) in both populations studied here. The broad-sense heritability estimates for CBSD in the present study were comparable to heritability estimates previously reported (Kayondo et al., 2018), ranging from 0.25 for CBSDRs to 0.61 for CBSD3s, from the genome wide association study, involving ~1,300 clones evaluated

across five sites in Uganda. These moderate to high broad-sense heritability estimates for CBSD, indicated that selection efficiency either through conventional or genomic selection would be high enough to achieve the desired genetic gains.

Fresh root weight and DMC, which are key traits for cassava production, had the lowest broad-sense heritability estimates. Root weight as a measure of yield is known to be polygenic and influenced by environment (Maria et al., 2002). According to Barandica et al. (2016), measures of yield and its components are best estimated at later stages of the cultivar selection pipelines, when plot sizes are larger than those in the current study. It is partly for this reason that harvest index has been proposed as an indirect measure of yield (Kawano et al., 1987). Indeed, in our study, harvest index had higher broad-sense heritability estimates than root weight, ranging between 0.2 and 0.44 compared with 0.14 and 0.30 for fresh root weight per plot. Thus, selecting on HI could be complementary to direct selection on fresh root yield, particularly at earlier stages of evaluation. The breeding program can then place heavier emphasis directly on root weight during later stages, when plot sizes are large enough for accurate assessment.

Broad-sense heritability estimates for DMC were particularly low (0.00-0.18), except for clones selected as parents of C_1. Clearly, our broad-sense heritability estimates for DMC were lower than the heritability estimate of 0.46 reported by Wolfe et al. (2017). In part, the low heritability estimates of DMC in the present study could be attributed to the effect of CBSD on the DMC. A previous study of CBSD effect on DMC reported significant differences between healthy roots and those with necrotic symptoms of CBSD (Nuwamanya et al., 2015). Often, infected roots with CBSD become necrotic; necrosis limits the quantity and quality of root samples used for DMC estimation via specific gravity method. Principally, specific gravity proposed by Kawano et al. (1987) uses 3-5 kg samples; in some cases, we used weights of less

than 3 kg for estimation of DMC. Furthermore, the adverse effect of CBSDRs on DMC is evident from the high negative genetic correlation (r = -0.64) observed in this study.

The SNP-based heritability estimates were generally lower than broad-sense heritability for most traits evaluated for C_1 clones. The precise reason for this seeming discrepancy between broad- and narrow-sense heritability estimates is not known. However, it has been shown both theoretically and with simulations on real data that one reason for the bias in heritability estimation using markers is variation in the amount of linkage disequilibrium (LD) between the markers and the causal loci. If the most important loci are in LD with many more markers than lesser causal loci, then the SNP-based heritability estimates can be upwardly biased. In contrast, if important causal loci are under-tagged by markers, the heritability estimates can be downwardly biased (Speed et al., 2012; De los Campos et al., 2015).

Thus, it is possible that high SNP-based heritability estimates observed for traits, such as CMD3s (h^2 = 0.81) in C_0 and DMC (h^2 = 0.79) in C_1 selected as parents, could be attributed to uneven LD between SNPs. Variation in LD has been previously reported in cultivated cassava (Bredeson et al., 2016), with notably low recombination rates observed in regions of introgression from a wild relative (*M. glazovii*) on chromosomes 1 and 4. These variations in LD patterns across the genome could therefore have led to over- or under-estimation of SNP-based heritability observed in the present study.

Estimates of phenotypic and genetic correlations among traits

Elsewhere, selection of parents at seedling stage for recombination has been reported to drastically shorten the breeding cycle of cassava for highly heritable traits (Ceballos et al., 2013). In this study, we observed a high genetic correlation (r_g = 0.70) between seedling and clonal CBSD root severities, suggesting the usefulness of seedling data in parental selection for recombination, training genomic selection models or in selection of clones for further evaluation, targeting cultivar release. Furthermore, CBSD has been reported to spread rapidly

in the last two decades in Africa (Hillocks et al., 2002; Alicai et al., 2007; Legg et al., 2011; Mulimbi et al., 2012) to cover countries other than the original CBSD-endemic coastal region of eastern Africa. The high genetic correlation observed between CBSD on seedlings and clonal stages would leverage pre-emptive breeding for CBSD in W. Africa through evaluation of botanical seeds from W. Africa in CBSD endemic areas.

We did not expect such low phenotypic correlation as observed for CMD assessments at seedling and clonal evaluation stages ($r_p \leq 0.12$), as it is a trait known to be highly heritable (Wolfe et al., 2016). Often, seedlings that are heavily infected with CMD (severity scores of >3) do not get cloned. Because we discarded seedlings that were highly infected with CMD at six MAP and did not collect leaf samples for DNA extraction from those, the overall genetic variance for CMD was decreased among the selected seedlings advanced to clonal stage. Also, some of the symptomless seedlings eventually succumbed to CMD at clonal evaluation. Scenarios of re-emergence of latent cassava mosaic virus have been observed previously from plants co-infected by two isolates, interacting in an antagonistic manner (Karthikeyan et al., 2016). One possible explanation would be that some seedlings had latent infection and eventually expressed CMD symptoms at clonal evaluation. A study by Ogbe et al. (2003) reported a low correlation between CMD symptom expression and virus titer, implying some genotypes harbored CMD-causing virus, without necessarily showing disease symptom, until such time as the virus population within the host plant had reached certain threshold to cause visible symptoms. Those phenomena could explain the low phenotypic correlation observed for seedling and clonal CMD data sets.

Examination of phenotypic and genetic correlations between disease severity and incidence for CBSD and CMD measured at three and six MAP revealed high positive genetic correlations ($r_g \geq 0.83$). Similar results were previously reported by Rwegasira and Rey (2012), where a phenotypic correlation of up to 0.98 was reported between foliar disease severity

scored at 3 MAP and 6 MAP of CBSD. While Rwegasira and Rey (2012) reported only phenotypic correlations for foliar CBSD severity scored at 3 MAP and 6 MAP, we report both phenotypic and genetic correlations for CBSD severity scored at 3 MAP and 6 MAP. In addition, we scored the foliar disease incidence at both 3 MAP and 6 MAP. These high correlations imply that data collected for disease incidence, especially on foliar plant health status would be sufficient and recommended, because scoring disease incidence is quicker and less subjective (absence or presence) than scoring disease severity on a wide scale (1-5). In contrast, there were very low phenotypic and genetic correlations between foliar CBSD symptoms (at three or six MAP) and CBSD root necrosis symptoms (at 12 MAP), which is consistent with earlier studies (Rwegasira and Rey, 2012).

Indeed, Nzuki et al. (2017) recently reported two QTL (located on chromosomes 1 and 12) to be significantly associated with CBSD root necrosis, and four other QTL (located on chromosomes 2, 4, 6, and 17) controlling foliar CBSD severity, suggesting some degree of independence in the genetic control of CBSD resistance. The high genetic correlation between foliar CBSD3s and CBSD6s in the current study implies the possibility of single and effective assessment of CBSD foliar symptoms at six MAP, permitting more efficient use of resources. Meanwhile, the high and positive phenotypic (r_p = 0.56) and genetic correlation (r_g = 0.41) between HI and root weight with large number of clones evaluated in the present study agree with the results of previous studies conducted in other breeding populations (Ojulong et al., 2010; Akinbo et al., 2012), suggesting HI could be used as a complementary trait for root weight per plot to select for fresh root yield, particularly at early stages of selection, when large number of clones are evaluated in smaller plots.

Population structure and level of inbreeding in C_0 and C_1 clones

We did not observe a distinct differentiation between C_0 and C_1 populations, which indicates that little or no genetic diversity was lost because of selection using genomic

predicted breeding values. Elsewhere, strong population stratification between the training set and the selection candidates has been reported to impact negatively on genomic prediction accuracies in oat and rice (Asoro et al., 2011; Grenier et al., 2016). In the present study, absence of population structures suggests appropriateness of using C_0 as a training population for genomic predictions of C_1 and subsequent selection of parents, using GEBVs. However, PC1 explained 49% of total genetic variation, suggesting sub-population structure, when considering the two populations jointly. Further examination of PC1 and PC2 marker scores across the 18 chromosomes of C_0 and C_1 populations revealed that markers with the strongest effects loaded on chromosomes 1 and 4 for PC1 and PC2, respectively (Figures. 3a and 3b). This finding corroborates with Bredeson et al. (2016), where chromosomes 1 and 4 were found to harbor large pieces of haplotype introgression from *Manihot glazovii* in many of the tropical *Manihot esculenta* (TME) and tropical *Manihot* selection (TMS) clones. These introgressions are believed to have occurred at the time of pioneer CMD and CBSD breeding at the Amani breeding station in Tanzania (Storey and Nichols, 1938). It also suffices to note that a significant number of the C_0 clones share ancestry with the TME and/or TMS lines that were introduced in Uganda between 1990s and early 2000s.

For both the C_0 and C_1 populations, the average of diagonal elements of K, as a measure of inbreeding coefficient (1+F) based on markers, was 0.904 and 0.708, respectively. These values should be interpreted to mean that the clones were less inbred than might be expected on the basis of the marker allele frequencies, i.e., the heterozygous marker genotypes were more frequent than expected under Hardy-Weinberg equilibrium. Cassava is known to suffer from inbreeding depression (Rojas et al., 2009; Kawuki et al., 2016; Ramu et al., 2017). Thus, selection among clones in establishing the C_0 population might have removed clones that were inbred. For the C_1 population, the priority given to between, rather than within, cluster crosses could also be expected to generate above average heterozygosity. Comparison of inbreeding

levels in C_0 and C_1 populations indicated less inbreeding in C_1 population than in C_0. As indicated, we think the crossing strategy we designed accounts for this observation. Evident in the current study was the better average performance of C_1 compared to C_0 population, an indication of overall genetic progress for most traits, which could be a result of less average inbreeding exhibited by C_1, as indicated by comparing the mean diagonals of the kinship matrix.

Conclusion

From the datasets presented, three major conclusions are drawn: First, seedling evaluation for CBSD, within limit, predicts CBSD clonal performance. This finding justifies selection for CBSD at the seedling stage; for this, use of both incidence and average root severity can suffice. Second, we observed moderate to high genetic correlations between foliar assessments made for CBSD and CMD at three and six months after planting. This finding justifies a single evaluation done at six months; such a strategy could significantly reduce costs associated with data collection in multi-location trials for foliar disease expression at an initial assessment of disease resistance. Third, selection on GEBVs did not erode the original genetic diversity and resulted in genetic progress for most traits as advances were made from C_0 to C_1. Based on these results, we do not expect genomic selection to cause rapid inbreeding, as breeding populations are moved from one cycle of genomic selection to the next.

Acknowledgments

This work was supported by the "Next Generation Cassava Breeding Project" through funds from the Bill and Melinda Gates Foundation and the Department for International Development of the United Kingdom. We thank the technical field staff of NaCRRI Cassava-breeding programme, who helped in the trial management and/or data collection (Joseph Orone, Charles

Majara, Gerald Adiga, and Vincent Kyaligonza). We thank the laboratory staff, particularly Francis Osingada and Jimmy Akano, for the support they provided during DNA extraction and quantification. We also thank Cornell University Institute of Genomic diversity (IGD) for carrying out the genotyping; Ramu Punna and Guillaume Jean Bauchet, from Cornell University, provided bioinformatics support

References

Akdemir, D., and U.G. Okeke. 2015. EMMREML: Fitting mixed models with known covariance structures, R Repository CRAN.

Akinbo, O., M. Labuschagne, and M. Fregene. 2012. Increased storage protein from interspecific F1 hybrids between cassava (*Manihot esculenta* Crantz) and its wild progenitor (*M. esculenta ssp. flabellifolia*). Euphytica 185: 303–311.

Alicai, T., C.A. Omongo, M.N. Maruthi, R.J. Hillocks, Y. Baguma, R. Kawuki, A. Bua, G.W. Otim-Nape, and J. Colvin. 2007. Re-emergence of cassava brown streak disease in Uganda. Plant Dis. 91(1): 24–29.

Asoro, F.G., M.A. Newell, W.D. Beavis, M.P. Scott, and J.-L. Jannink. 2011. Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. Plant Genome J. 4: 132.

Barandica, O.J., J.C. Pérez, J.I. Lenis, F. Calle, N. Morante, L. Pino, C.H. Hershey, H. Ceballos, and D.M.O. Sullivan. 2016. Cassava breeding II : Phenotypic correlations through the different stages of selection. Front. Plant Sci. 7: 1–11.

Bernardo, R. 2003. Breeding for quantitative traits in plants. Stemma Press, Woodbury, Minnesota.

Bhat, J.A., S. Ali, R.K. Salgotra, Z.A. Mir, S. Dutta, V. Jadon, A. Tyagi, M. Mushtaq, N.

Jain, P.K. Singh, G.P. Singh, and K. V. Prabhu. 2016. Genomic selection in the era of next generation sequencing for complex traits in plant breeding. Front. Genet. 7: 1–11.

Bredeson, J. V, J.B. Lyons, S.E. Prochnik, G.A. Wu, C.M. Ha, E. Edsinger-Gonzales, J. Grimwood, J. Schmutz, I.Y. Rabbi, C. Egesi, P. Nauluvula, V. Lebot, J. Ndunguru, G. Mkamilo, R.S. Bart, T.L. Setter, R.M. Gleadow, P. Kulakow, M.E. Ferguson, S. Rounsley, and D.S. Rokhsar. 2016. Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. Nat. Biotechnol. 34: 562–570.

Burns, A., R. Gleadow, J. Cliff, A. Zacarias, and T. Cavagnaro. 2010. Cassava: The drought, war and famine crop in a changing world. Sustainability 2: 3572–3607.

Ceballos, H., N. Morante, T. Sánchez, D. Ortiz, I. Aragón, A.L. Chávez, M. Pizarro, F. Calle, and D. Dufour. 2013. Rapid cycling recurrent selection for increased carotenoids content in cassava roots. Crop Sci. 53: 2342-2351.

Ceballos, H., J.C. Pérez, O.J. Barandica, J.I. Lenis, N. Morante, F. Calle, L. Pino, and C.H. Hershey. 2016. Cassava breeding I: The value of breeding value. Front. Plant Sci. 7: 1–12.

Ceron-Rojas, J.J., J. Crossa, V.N. Arief, K. Basford, J. Rutkoski, D. Jarquín, G. Alvarado, Y. Beyene, K. Semagn, and I. DeLacy. 2015. A genomic selection index applied to simulated and real data. Genes|Genomes|Genetics 5(10): 2155–2164.

Edgerton, M.D. 2009. Increasing crop productivity to meet global needs for feed, food, and fuel. Plant Physiol. 149: 7–13.

Egesi, C.N., F.O. Ogbe, M. Akoroda, P. Ilona, and A. Dixon. 2007. Resistance profile of improved cassava germplasm to cassava mosaic disease in Nigeria. Euphytica 155: 215–224.

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

Endelman, J.B. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome J. 4: 250.

Endelman, J.B., and J.-L. Jannink. 2012. Shrinkage estimation of the realized relationship matrix. Genes|Genomes|Genetics 2: 1405–1413.

Esuma, W., R.S. Kawuki, L. Herselman, and M.T. Labuschagne. 2016. Diallel analysis of provitamin A carotenoid and dry matter content in cassava (*Manihot esculenta* Crantz). Breed. Sci. 66: 627–635.

Fischer, R.A., and G.O. Edmeades. 2010. Breeding and cereal yield progress. Crop Sci. 50: 85–98.

Garrick, D.J., J.F. Taylor, and R.L. Fernando. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet. Sel. Evol. 41: 55.

Grenier, C., T.V. Cao, Y. Ospina, C. Quintero, M.H. Châtel, J. Tohme, B. Courtois, and N. Ahmadi. 2016. Correction: Accuracy of genomic selection in a rice synthetic population developed for recurrent selection breeding. PLoS ONE 11(5): e0154976.

Hamblin, M.T., and I.Y. Rabbi. 2014. The effects of restriction-enzyme choice on properties of genotyping-by-sequencing libraries: A study in Cassava (*Manihot esculenta*). Crop Sci. 54: 2603–2608.

Hillocks, R.J., and M.N. Maruthi. 2015. Post-harvest impact of cassava brown streak disease in four countries in eastern Africa. Food Chain 5: 116–122.

Hillocks, R.J., and J.M. Thresh. 2000. Cassava mosaic and cassava brown streak virus diseases in Africa. Roots 7(1) Special issue: 1–8.

Hillocks, R.J., J.M. Thresh, J. Tomas, M. Botao, R. Macia, and R. Zavier. 2002. Cassava brown streak disease in northern Mozambique. Int. J. Pest Manag. 48: 178–181.

Iglesias, C., and D. J Jennings. 2002. Cassava: biology, production and utilization. Hillocks, R.J., J.M. Thresh (eds), Natural Resources Institute, University of Greenwich, UK, A Bellotti, Centro Internacional de Agricultura Tropical (CIAT), Cali, Colombia, pp 149.

Iglesias, C.A., J.C. Pe, and A.G.O. Dixon. 2004. Cassava breeding : opportunities and challenges. Plant Mol. Biol. 56: 503–516.

IITA. 1990. Cassava in tropical Africa: A reference manual. International Institute of Tropical Agriculture, Abadan, Nigeria.

Jannink, J.-L., A.J. Lorenz, and H. Iwata. 2010. Genomic selection in plant breeding: from theory to practice. Brief. Funct. Genomics 9: 166–177.

Karthikeyan, C., B.L. Patil, B.K. Borah, T.R. Resmi, S. Turco, M.M. Pooggin, T. Hohn, and K. Veluthambi. 2016. Emergence of a latent Indian cassava mosaic virus from cassava which recovered from infection by a non-persistent Sri Lankan cassava mosaic virus. Viruses 8(10): 264-278.

Kawano, K. 1980. Cassava. In; Fehr W. R and Hadley H. H (eds). Hybridization of Crops plants. ASA, CSSA, Madison, Wisconsin. pp. 225–233.

Kawano, K., W.M.G. Fukuda, and U. Cenpukdee. 1987. Genetic and environmental effects on dry matter content of cassava root. Crop Sci. 27: 69-74.

Kaweesi, T., R. Kawuki, V. Kyaligonza, Y. Baguma, G. Tusiime, and M.E. Ferguson. 2014. Field evaluation of selected cassava genotypes for cassava brown streak disease based on symptom expression and virus load Field evaluation of selected cassava genotypes for cassava brown streak disease based on symptom expression and virus load. Virol. J. 2: 11–216.

Kawuki, R.S., T. Kaweesi, W. Esuma, A. Pariyo, I.S. Kayondo, A. Ozimati, V. Kyaligonza, A. Abaca, J. Orone, R. Tumuhimbise, E. Nuwamanya, P. Abidrabo, T. Amuge, E. Ogwok, G. Okao, H. Wagaba, G. Adiga, T. Alicai, C. Omongo, A. Bua, M. Ferguson, E. Kanju, and Y. Baguma. 2016. Eleven years of breeding efforts to combat cassava brown streak disease. Breed. Sci. 66: 560–571.

Kayondo, S.I., D.P. Del Carpio, R. Lozano, A. Ozimati, M. Wolfe, Y. Baguma, V. Gracen, S. Offei, M. Ferguson, R. Kawuki, and J.L. Jannink. 2018. Genome-wide association mapping and genomic prediction for CBSD resistance in *Manihot esculenta*. Sci. Rep. 8: 1–11.

Legg, J.P., S.C. Jeremiah, H.M. Obiero, M.N. Maruthi, I. Ndyetabula, G. Okao-Okuja, H. Bouwmeester, S. Bigirimana, W. Tata-Hangy, G. Gashaka, G. Mkamilo, T. Alicai, and P. Lava Kumar. 2011. Comparing the regional epidemiology of the cassava mosaic and cassava brown streak virus pandemics in Africa. Virus Res. 159: 161–170.

Lloyd, S.P. 1982. Least squares quantization in PCM. IEEE Trans. Inf. Theory 28: 129–137.

De los Campos, G., D. Sorensen, and D. Gianola. 2015. Genomic heritability: What is it? PLoS Genet 11(5): e1005048.

Lynch, M., and B. Walsh. 1998. Genetics and Analysis of Quantitative Traits, By Michael Lynch and Bruce Walsh. Sunderland, MA: Sinauer Associates, Inc., 1998. pp. 980

Maria, W., G. Fukuda, S. De Oliveira, and C. Iglesias. 2002. Cassava breeding. Crop Breed. Appl. Biotechnol. 2: 617–637.

Meuwissen, T. H. E. , B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense markers maps. Genetics 157: 1819–1829.

Mezzalira, I., C.J. Costa, E.A. Vieira, J. Freitas, M.. Silva, M.L. Denke, and K. Nascimento.

2013. Pre-germination treatments and storage of cassava seeds and their correlation with emergence of seedlings. J. Seed Sci. 35: 113–118.

Mulimbi, W., X. Phemba, B. Assumani, P. Kasereka, S. Muyisa, H. Ugentho, R. Reeder, J.P. Legg, L. Laurenson, R. Weekes, and F.E. Thom. 2012. First report of Ugandan cassava brown streak virus on cassava in Democratic Republic of Congo. New Dis. Reports 26: 11.

Muranty, H., M. Troggio, I. Ben Sadok, M. Al Rifaï, A. Auwerkerken, E. Banchi, R. Velasco, P. Stevanato, W.E. van de Weg, M. Di Guardo, S. Kumar, F. Laurens, and M.C.A.M. Bink. 2015. Accuracy and responses of genomic selection on key traits in apple breeding. Hortic. Res. 2: 15060.

Njoku, D.N., V.E. Gracen, S.K. Offei, I.K. Asante, C.N. Egesi, P. Kulakow, and H. Ceballos. 2015. Parent-offspring regression analysis for total carotenoids and some agronomic traits in cassava. Euphytica 206: 657–666.

Nuwamanya, E., B. Yona, A. Evans, A. Sharon, and A. Titus. 2015. Effect of cassava brown streak disease (CBSD) on cassava (*Manihot esculenta* Crantz) root storage components , starch quantities and starch quality properties. Int. J. Plant Physiol. Biochem. 7: 12–22.

Nzuki, I., M.S. Katari, J. V. Bredeson, E. Masumba, F. Kapinga, K. Salum, G.S. Mkamilo, T. Shah, J.B. Lyons, D.S. Rokhsar, S. Rounsley, A.A. Myburg, and M.E. Ferguson. 2017. QTL mapping for pest and disease resistance in cassava and coincidence of some QTL with introgression regions derived from *Manihot glaziovii*. Front. Plant Sci. 8: 1–15.

Ogbe, F.O., G.I. Atiri, A.G.O. Dixon, and G. Thottappilly. 2003. Symptom severity of cassava mosaic disease in relation to concentration of African cassava mosaic virus in different cassava genotypes. Plant Pathol. 52(1): 84–91.

Ojulong, H.F., M.T. Labuschagne, L. Herselman, and M. Fregene. 2010. Yield traits as

selection indices in seedling populations of cassava. Crop Breed. Appl. Biotechnol. 10: 191–196.

Olsen, K.M., and B.A. Schaal. 1999. Evidence on the origin of cassava: Phylogeography of *Manihot esculenta*. Proc. Natl. Acad. Sci. 96(May): 5586–5591.

Pariyo, A., Y. Baguma, T. Alicai, R. Kawuki, E. Kanju, A. Bua, C.A. Omongo, P. Gibson, D.S. Osiru, and D. Mpairwe. 2015. Stability of resistance to cassava brown streak disease in major agro-ecologies of Uganda. J. Plant Breed. Crop Sci. 7: 67–78.

R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. R Found. Stat. Comput. Vienna, Austria (ISBN 3-900051-07-0): 900051.

Ramu, P., W. Esuma, R. Kawuki, I.Y. Rabbi, C. Egesi, J. V. Bredeson, R.S. Bart, J. Verma, E.S. Buckler, and F. Lu. 2017. Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. Nature Genetics 49: 959–963.

Rojas, M.C., J.C. Pérez, H. Ceballos, D. Baena, N. Morante, and F. Calle. 2009. Analysis of inbreeding depression in eight S1 cassava families. Crop Sci. 49: 543–548.

Rwegasira, G.M., and C. M. E. Rey. 2012. Response of selected cassava varieties to the incidence and severity of cassava brown streak disease in Tanzania. J. Agric. Sci. 4: 237.

Speed, D., G. Hemani, M.R. Johnson, and D.J. Balding. 2012. Improved heritability estimation from genome-wide SNPs. Am. J. Hum. Genet. 91: 1011–1021.

Storey, H.H., and R.F.W. Nichols. 1938. Studies of the mosaic disease of cassava. Ann. Appl. Biol. 25: 790–806.

Thresh, J.M., and R.J. Cooter. 2005. Strategies for controlling cassava mosaic virus disease in Africa. Plant Pathol. 54: 587–614.

Wolfe, M.D., D.P. Del Carpio, O. Alabi, L.C. Ezenwaka, U.N. Ikeogu, I.S. Kayondo, R. Lozano, U.G. Okeke, A.A. Ozimati, E. Williams, C. Egesi, R.S. Kawuki, P. Kulakow, I.Y. Rabbi, and J.-L. Jannink. 2017. Prospects for genomic selection in cassava breeding. Plant Genome 10: 1–19.

Wolfe, M.D., I.Y. Rabbi, C. Egesi, M. Hamblin, R. Kawuki, P. Kulakow, R. Lozano, D.P. Del Carpio, P. Ramu, and J.-L. Jannink. 2016. Genome-wide association and prediction reveals genetic architecture of cassava mosaic disease resistance and prospects for rapid genetic improvement. Plant Genome 9: 342–356.

CHAPTER 3

INCORPORATING GENOME-WIDE MARKERS AND WEATHER VARIABLES IN GENOTYPE-BY-ENVIRONMENT INTERACTION ANALYSES

Abstract

Genotype-by-environment (G x E) interaction is a reality that scientists deal with when developing new and better varieties. Accordingly, a number of approaches have been suggested to undertake G x E analyses, ranging from classical analysis of variance in fixed effects model to linear mixed models. With increase in scale of both phenotypic and genetic data, coupled with environmental data, prediction of environment-specific response of genotypes is gaining importance. We used phenotypic and genotypic data for ~150 clones and five checks evaluated in 31 environments (location-season-year combination), alongside four weather variables, to define mega environments. Further, we used this dataset to address different prediction problems faced in breeding: (i) predicting unobserved genotypes across environments, (ii) predicting unobserved genotypes in unobserved environments and (iii) making predictions for unobserved environments. Phenotypic data of the five-checks clustered environments by seasons. Our prediction accuracies for unobserved genotypes (cross validation 1, CV1, mimicking new crosses) across environments were moderate to high for cassava brown streak disease (CBSD), dry matter content (DMC) and harvest index (HI), ranging from 0.45 for CBSD root severity to 0.61 for CBSD foliar severity among the models tested, whereas low predictive abilities were observed for shoot and root weight per plot (r = 0.04 to 0.08) across models. Similar results across traits and models were observed for the prediction scheme of unobserved genotypes in unobserved environments (cross validation 2, CV2). Highest prediction accuracies were recorded for leave-one-environment-out cross validation (CV3), which varied from 0.74 for shoot weight to 0.95 for HI. From this study, we established that

CBSD, HI and DMC can be predicted with reasonable accuracies under different scenarios that mimic real problems encountered in cassava breeding.

Manuscript written: Alfred Ozimati, Robert Kawuki, Williams Esuma, Deniz Akdemir, Marnin Wolfe, and Jean-Luc Jannink. 2018. Genomic-wide Markers and Weather Variables to Predict Response of Cassava Genotypes in a Multi-locational Trials

Abbreviations

G x E; genotype-by-environment interaction; MET, multi-environment trials; CV, cross validation; TPOE, target population of environments; CET, clonal yield trial; PYT preliminary yield trial; TP, training population; GEBV, genomic estimated breeding value; G-BLUP, genomic best linear unbiased predictor; GS, genomic selection; SNP, single nucleotide polymorphism.

Introduction

The target of a plant breeding program is to release varieties that perform consistently better than the existing varieties grown by farmers (Bernardo, 2003). At times, the response of genotypes across environments is not consistent due to the diversity of conditions in farmers' fields, a phenomenon known as genotype-by-environment (G x E) interaction. To mitigate the impact of G x E and develop varieties with wide adaptation, plant breeders conduct extensive multi-environment trials to accurately assess the performance of the genotypes in the target population of environments (TPOE) (Van Eeuwijk et al., 2016). The comprehensive genotype evaluations are done because the phenotypic expression of quantitative traits is not completely under genetic control, as it is known that the metabolic and developmental pathways for most quantitative traits are influenced by aspects of environment (Lynch and Walsh, 1998; Jarquín et al., 2014)

In cassava, evaluation of clones at early stages of the breeding pipeline (i.e., the clonal evaluation trial, CET, or preliminary yield trial, PYT) is done on the basis of performance

assessment carried out in less than three locations. Evaluation can only be done in a few locations because of the large number of entries that need to be evaluated and the limited amount of stem cuttings that are available for each clone at those evaluation stages (Ceballos et al., 2007). The limited number of evaluation sites does not provide adequate assessment of the G x E for the clones selected for further evaluations.

One strategy breeders use to reduce the impact of G x E is to partition the TPOE into smaller homogenous groups of environments that may have similar growing conditions such as soil type, temperature, precipitation or biotic stresses (Bernardo, 2003). Therefore, one objective of this study was to delineate the mega environments for cassava production in Uganda.

Previous G x E interaction studies conducted in cassava revealed significant G x E interactions for important agronomic traits such as fresh root yield, harvest index, as well as quality attributes such dry matter and total carotene content (Tumuhimbise et al., 2014; Pariyo et al., 2015; Esuma et al., 2016). Similarly, other studies have demonstrated significant G x E interaction for cassava brown streak disease (CBSD), which is caused by two virus species, the cassava brown streak virus, CBSV and the Uganda variant, UCBSV. This disease is currently prevalent in the Eastern, Central and Southern regions of Sub-Saharan Africa (Pariyo et al., 2015; Mtunguja et al., 2016; Masinde et al., 2018). The early G x E interaction studies used analytic tools such as the graphical AMMI and GGE biplots to model G x E for hypothesis testing in fixed effect analysis of variance (ANOVA) (Tumuhimbise et al., 2014; Esuma et al., 2016; Mtunguja et al., 2016; Masinde et al., 2018). Although graphical presentation of G x E analyses provides a quick visualization (Yan et al., 2007), these analyses treat G x E as a black-box, making their biological interpretations difficult (Malosetti et al., 2013; Jarquín et al., 2014).

Recently, plant breeding has undergone a revolution due to an increase in the scale of both phenotypic and genetic data generation, further boosted by increases in computational efficiency in mixed linear models (Cooper et al. 2014). The emergence of such big data and the techniques necessary to analyze them have enabled the application of a new breeding and selection method known as genomic selection (Meuwissen et al. 2001). In genomic selection (GS), the breeding value of new individuals, not yet observed in the field, can be predicted at early stages based on their genetic relationships to a phenotyped and genotyped calibration set known as the training population (TP) (Hayes et al., 2009). Most applications of GS, especially early on, focused on genetic main-effects, that is on selecting lines that performed well overall. They did not model G x E directly, or attempt to predict environment-specific breeding values. However, recently a number studies have recognized the need to account for G x E in the genomic prediction framework (Burgueño et al., 2012; Heslot et al., 2014; Jarquín et al., 2014).

As recently as a decade ago, cassava lagged behind many other crops species such as wheat, rice, and maize in terms of genomic resources. Since then a large amount of attention has been given to cassava, leading to rapid development of genomic resources (Prochnik et al., 2012), which have further enabled access to relatively cheap genotyping technologies like genotyping-by-sequence (GBS) (Elshire et al., 2011; Hamblin and Rabbi, 2014). Two recent publications have leveraged GBS data to assess the accuracies of genomic prediction in light of G x E for important agronomic and disease traits in cassava (Ly et al., 2013; Okeke et al., 2017). However, both papers model G x E using the location of phenotyping trials as proxy for environmental variation. According to Bernardo, (2003), additional information from external environmental factors such as rainfall, temperature, solar radiation and soil parameters measured in routine performance trials could provide relevant biological descriptions and understanding of G x E.

Based on this need to model G x E in a GS framework, a number of studies have tested prediction models that incorporate environmental covariates, allowing information sharing from environments of interest (Heslot et al., 2014; Jarquín et al., 2014; Lado et al., 2016; Ly et al., 2018). These studies indicated increases in genomic prediction accuracies are possible when G x E is modelled explicitly using environmental covariates because they improve information sharing among environments. Given these benefits, the use of environmental covariates in genomic prediction models could be worthwhile in cassava. Genomic prediction of G x E in cassava is expected to permit more optimal resource allocation to boost genetic gains without significantly increasing breeding costs.

There are a number of common breeding problems such as the prediction of unobserved genotypes (newly generated crosses) or environments (environments not previously tested), where genomic prediction with environmental covariates might be beneficial in cassava. In order to assess the prediction accuracy possible in these scenarios, cross-validation schemes are typically set-up so as to mimic real prediction scenarios (De Los Campos et al., 2009; Crossa et al., 2010; Burgueño et al., 2012). Most studies that incorporate G x E into genomic prediction have used two basic cross-validation schemes (Burgueño et al., 2012): (i) predicting the performance of unobserved genotypes across environments (CV1) and (ii) predicting the performance of unobserved genotypes in unobserved environments (CV2). Another important scenario is the prediction of environments (i.e., site-season-year combinations) that are not included in the training population set of testing environments. This scenario is evaluated using leave-one-environment-out cross-validation, as described in a recent study (Jarquín et al. 2017).

The main objective in this study was to leverage environmental data to improve prediction of key cassava traits in the presence of G x E. Our specific aims were to: (i) identify the mega environments for cassava breeding in Uganda, and (ii) test genomic prediction for

unobserved genotypes across sites, unobserved genotypes in unobserved environments, and prediction of fully-unobserved environments (leave-one-environment-out).

Materials and Methods

Genetic materials and field evaluations

A total of 150 clones that were part of the initial 427 genotypes comprising the training population for implementation of genomic selection at the National Crops Resources Research Institute (NaCRRI) in Uganda were selected for this experiment (Ozimati et al., 2018). Of the 150 clones, 100 were initially selected as progenitors to generate the first cycle (C_1) of GS progenies. The 100 parents were chosen based on a selection index combining standardized genomic estimated breeding values (GEBVs) for cassava brown streak disease (CBSD), root weight (RTWT) per plot, harvest index (HI), and dry matter content (DMC) as described in chapter two. Additionally, 50 random clones were selected from the remaining training population to make a total of 150 clones. These clones were planted at 10 sites, chosen to represent the major cassava production and consumption areas in Uganda. The sites included Namulonge, Kigumba and Mityana in the Central, Kasese in the West, Kamuli, Pallisa, Serere and Kaberamaido in the East, and Lira and Arua in the Northern part of Uganda (Figure 3. 1).

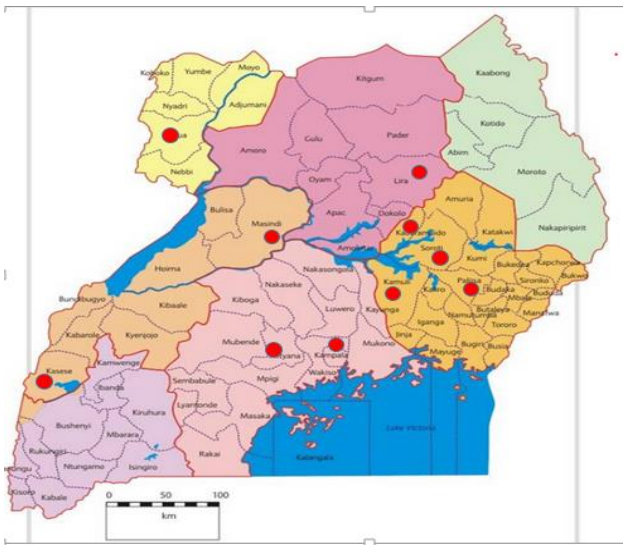


Figure 3. 2: Map of Uganda indicating the trial sites

The experiments were set in an augmented design with five common checks (UG110008, UG110014, UG110015, UG110016 and UG110017). Each check was replicated 5-6 times and was represented in every block (Federer, 1954). We collected data on CBSD and cassava mosaic disease (CMD) foliar symptoms scored at three and six months after planting (3 and 6 MAP). For both diseases, we scored severity on the standard scale of 1 (low severity, no disease) to 5 (high severity, severe disease) (IITA, 1990; Hillocks and Thresh, 2000). At harvest, all roots in a plot were pooled and assessed individually for CBSD necrosis. Each root was cut transversely, and the cross-sections were scored for necrotic symptoms on a scale of 1-5. Additionally, we collected data on the fresh root and shoot weights per plot from which HI was calculated as a ratio of root weight to total biomass. To determine the DMC, we used the specific gravity method as described by Kawano et al. (1987).

Uganda typically receives bimodal rainfall and the farmers plant cassava following the onset of rains in the two rainy seasons. We established the experiments to cover these two planting seasons, with the first season rains occurring in the months of Feb-May and the second season rains occurring from the months of Aug-Dec. We carried out the evaluation at the ten sites previously mentioned, following the two planting seasons (Feb-May and Aug-Dec) for a period of two years (2015-2017). Our initial target number of trials was 40 (all location-season-year combinations). However, the loss of nine field trials due to severe drought resulted in 31 field trials with complete data.

In addition to the phenotypic data, we collected weather information using data loggers (HOBO® Pro v2: www.onsetcomp.com) that were installed at each of the 10 experimental sites to record rainfall (mm), temperature (˚C), solar radiation (lux) and relative humidity (%) for the two years of the experiment. The weather loggers were set to record the four variables every 15 min. The weather data were summarized as monthly averages to be used subsequently as environmental covariates in downstream prediction analyses.

DNA extraction and genotyping

For genotyping, DNA was extracted from approximately 100 mg of fresh young leaves from each of the 150 clones. All extractions were done using QIAGEN DNeasy extraction kits and DNA was quantified to ensure the required concentrations for sequencing were obtained. The DNA samples were genotyped using the genotyping-by-sequencing (GBS) method described by Elshire et al. (2011). Calling of SNP, filtering and imputation method have been described previously (Hamblin and Rabbi, 2014; Wolfe et al., 2016; Wolfe et al. 2017). Ultimately, we had a total of 25,383 single nucleotide polymorphic (SNP) markers with a minor allele frequency (MAF) ≥ 0.01 for the subsequent prediction analyses.

Statistical Analyses

Delineating the mega-environments

We used a cluster analysis approach to classify the environments. First, we clustered environments based on the phenotypic data for nine traits (CBSD3s, CMD3s, CBSD6s, CMD6s, CBSDRs, SHTW, RTWT, HI, and DMC) from the five checks, giving a total of 45 variables for clustering. We only used the data from the checks because they were the only clones that were observed at least once in each of the 31 environments (Figure S3. 1). Second, we clustered the environments using the four weather variables described above, each averaged over 12 month-long periods, giving 48 variables for clustering. For both the phenotypic and weather variables used for clustering the environments, we normalized the data by centering and scaling each variable. In each case, we constructed a distance matrix based on the normalized data using the *dist* function in R. Further for visualization, we used the hierarchical clustering function *hclust* to generate a dendrogram using the "Ward D" method (Murtagh and Legendre, 2011).

To compare the relationship between clustering based on phenotypic data to the clustering based on weather variables, we extracted the off-diagonal elements of the matrix of

among-environment Euclidean distances. We then computed the Pearson's correlation (*r*) between the phenotypic-based and the weather variable-based distance matrix.

Testing the relevance of genotype-by-environment interactions

To test the importance of G x E for the traits investigated, two mixed linear models, one with the genotype-by-environment interaction term (full model) and the second model without genotype-by-environment (reduced model) were fitted, using the function *lmer* in the R package *lme4* as follows.

$y_{ijk} = \mu + G_i + E_j + GE_{ij} + \varepsilon_{ijk}$…….....Full model

$y_{ijk} = \mu + G_i + E_j + \varepsilon_{ijk}$…….....Reduced model

Where, y_{ijk} was the response of the k^{th} replicate for the i^{th} genotype in the j^{th} environment, μ represented the fixed trial mean, G was the vector of random genotype effects, E was the vector of random environment effects (location-season-year combination), while GE represented a vector of random a two-way genotype-by-environment interaction effects. In order to assess the relevance of G x E, a chi-square test comparing the full model to the reduced model (no G x E term), using the deviance values was performed for each trait with the *anova* function. For all subsequent analyses, we focused only on traits for which G x E was found to be important at this stage.

We next tested the importance of the each of the 48 (4 covariates x 12 months) environmental covariates in terms of their ability to account for phenotypic variation between clones (G x E variation). For each trait and each covariate, we fitted the mixed linear model with a fixed trial and main environment effects, while the environment covariate nested within genotype (reaction-norm, G x E) term was considered random as below;

$$y_{ijk} = \mu + E_j + b_i W_j + G_i + e_{ijk}$$

Where, y_{ijk} was the response of k^{th} replicate for the i^{th} genotype in the j^{th} environment, μ and E represented the fixed trial mean and environment main effects respectively, b_i is the sensitivity

of the i^{th} genotype to environmental covariate W, and G_i is the main genotypic effect for the i^{th} genotype. Both b_i and G_i as well as e (the residual term) were treated as random effects. Since our trials were planted at different months (capturing the two seasons), we did not analyze data across trials using calendar month-based averages. Instead, we aligned the monthly averages to the month-of-planting for each environment such that the first month average for each weather variable was the planting month for that trial and the twelfth month was the month of harvest. This resulted in a total of 48 weather variables tested for their importance in accounting for G x E. For visualization, we plotted the variance-estimates for the reaction-norm term, b_i, across the genotypes (Figure S3.2).

Furthermore, we conducted a principal component analysis (PCA) on the 48 calculated weather variables recorded across the 31 environments, and plotted the loading of the first and second principal components (PCs) against each weather variable (Figure S3.3).

Genomic prediction models

The analyses below involved an environment covariate-based covariance matrix (Ω) among environments, computed from 48 standardized weather variables using *tcrossproduct* function built in R (R Development Core Team, 2008). While the SNP-based relationship matrix (**G**) among clones was computed using *A.mat* function in rrBLUP package (Endelman, 2011). These matrices were used as follows in the three genomic prediction models tested.

(i) The "NoGxECovariates" model, with genotypic and environmental main effects, with a random environmental main-effect normally distributed with variance-covariance matrix (Ω) and a standard random genotype main-effect with variance-covariance matrix (**G**):

$y_{ijk} = \mu + w_i + g_j + \varepsilon_{ijk}$.....(NoGxECovariates Model)

Where, y_{ijk} was the response of replicate k of the j^{th} genotype in the i^{th} environment, μ was the fixed grand mean, w was the random effect of the i^{th} environment, assuming $w_i \sim N(0, \Omega\sigma^2_w)$ with σ^2_w representing the variance due to environmental effects, and g was the random

genotype effect, assuming $g_j \sim N(0, \mathbf{G}\sigma^2_g)$ with σ^2_g representing the variance due to genotypic effects, whereas ε was the random model residual effect, assumed to be normally distributed as, $\varepsilon_{ijk}^{IID} \sim N(0, \sigma^2_\varepsilon)$ with σ^2_ε as the residual variance. The covariance matrices **Ω** and **G** permit the borrowing of information between environments and between genotypes, respectively.

(ii) The "GxENoCovariates" model, with genotypic main-effect as in (i), but with environment main-effects independently and identically distributed (IID), equivalent to using an identity matrix instead of **Ω**, and including a random G x E interaction term:

$y_{ijk} = \mu + E_i + g_j + gE_{ij} + \varepsilon_{ijk}$ ..... (GxENoCovariates Model)

Where, y_{ijk}, μ and g_j were as described in the model above (NoGxECovariates Model), E was the random effect of the i^{th} environment, assuming $E_i^{IID} \sim N(0, I\sigma^2_E)$ with I and σ^2_E representing the identity matrix and variance due to the environments, respectively. While gE represented the random genotype-by-environment interaction (G x E) effects, assuming $gE_{ij} \sim N(0, I_E \otimes \mathbf{G}\, \sigma^2_{gE})$, with $I_E \otimes \mathbf{G}$ being the Kronecker product of the environment identity matrix and SNP-based realized relationship matrix (**G**). Similar to the NoGxECovariates Model, ε was the random model residual effect assumed to be normally distributed as, $\varepsilon_{ijk}^{IID} \sim N(0, \sigma^2_\varepsilon)$ with σ^2_ε as the residual variance. Because the environments were considered to be IID, there was no sharing of information between environments in this model.

(iii) The "GxECovariate" model, with G x E variance-covariance matrix derived from both **G** and Ω:

$y_{ijk} = \mu + w_i + g_j + gw_{ij} + \varepsilon_{ijk}$ ..... (GxECovariates Model)

Where; y_{ijk}, μ, w_i, g_j and ε_{ijk} terms were modeled as described previously in the two models. However, the genotype-by-environment random term, gw was modeled explicitly using marker and environmental covariance matrices, assuming $gw_{ij} \sim N(0, \Omega \otimes \mathbf{G}\, \sigma^2_{gw})$ with Ω⊗G being the Kronecker product of the environment variance-covariance matrix (Ω) constructed from the environmental covariates and SNP-based realized relationship matrix (**G**). Since, the model

(GxECovariates) takes into account both the environment and genotype variance-covariance structure, information sharing explicitly via environmental covariates and genomic markers is made possible.

Cross-validation scenarios to assess genomic prediction accuracy of G x E models

The first two cross-validation scenarios we tested were previously described by Jarquín et al., (2014). In the first (CV1), genomic predictions were made for genotypes that had not been evaluated in any environments. This prediction scenario mimics predicting the performance of newly developed genotypes (crosses) or introductions into the breeding program. The second cross-validation scenario (CV2) involved predicting the performance of never-before tested individuals in environments that had provided no previous data as a strategy to assess ability to predict the performances of a new clone in a new environment.

In order to carry out the cross-validation for CV1 and CV2 scenarios, we used 130 clones, a subset of the 150 clones that had both marker and phenotype data. For CV1, we divided the clone list into five non-overlapping sets (folds). We then fitted a single step multi-kernel genomic prediction model to predict each fold, with four folds as the training set (phenotypes in model), and one fold as the validation set (no phenotypes). We used the function *emmremlMultiKernel* in the EMMREML R package (Akdemir and Okeke 2015). For CV1, all 31 environments were used in this single step analysis.

The CV2 scenario used the same five folds of clones as CV1. In addition, the 31 environments were split into five groups of about six environments. Validation data were taken from one group, while the remaining four groups were used for training the model. For each of the five-folds across clones, the analysis was repeated for each of the five groups of environments. The group was excluded from the training data, and the single step analysis was performed on the remaining 25 environments. Predictions from those environments were used to predict performance in each of the six excluded environments from the group.

To compute prediction accuracy per trait for the models with a G x E term, we first fitted single-step genomic prediction model using *emmremMultiKernel* function, from which the genotype main-effect (g_i) and the G x E (wg_{ij}) best linear unbiased predictors (BLUPs) were extracted. To obtain the observed estimated genetic value, we fitted a mixed linear model in *lme4* package as shown below;

$y_{ijkl} = \mu + G_i + E_j + B/E_{lj} + GE_{ij} + \varepsilon_{ijkl}$

Where the grand mean μ and the main effect of the j^{th} environment (E) (location-season-year combination) were considered fixed, while the i^{th} genotype (G) main effect for k^{th} replicate, the l^{th} incomplete block nested within the j^{th} environment (E), the interaction effect of the i^{th} genotype (G) by j^{th} environment (GE) and residual term (ε) were considered random, from which the genotype main effect (G) and the interaction effect (GE) BLUPs were also extracted.

Finally, we calculated the prediction accuracies as correlation between the sum of main genotype effects and genotype-by-interaction BLUPs extracted from *emmremMultiKernel* and *lme4* models e.g., for the GxECovariates model, $r = cor(g_i + gw_{ij}, G_i + GE_{ij})$, where *r* is the Pearson's correlation coefficient that measures the model prediction ability. For the genomic prediction model without G x E, accuracy per trait was calculated as Pearson's correlation between the genotypic main effects (g_i) from *emmremMultiKernel*, and sum of genotypic main (G) and interaction (GE) term BLUPs from *lme4* models i.e. $cor(g_i, G_i + GE_{ij})$. The five-fold cross-validation scheme was replicated five times, resulting in 25 prediction accuracies per trait for both CV1 and CV2.

The last prediction problem (CV3) did not involve random cross-validation per-se. This was a scenario predicting the performance of clones in previously unobserved environments. In a real breeding program, this prediction problems mimics predicting the performance of an environment a breeder has never used for evaluation, potentially aided by environmental covariates measured in the new environment, assuming it is part of the target population

environments for breeding. For this scenario, each prediction was trained on the data in 30 environments and predictions were validated against the data from each of the 31 test environments (one environment excluded from training to constitute the validation set). Prediction ability per trait was calculated for the G x E model, similar to the approach described above for CV1 and CV2 schemes.

Results

Characterization of the environments

Clustering of the environments using the phenotypic and monthly-average weather data are presented in Figure 3.3 and 3.4, respectively. Based on grouping of the environments from the phenotypic data of the five checks, the 31 environments separated into two groups at a tree height of 4, with the first cluster comprising 18 environments of which 13 were from the first growing season (Feb-May rains), which we refer to as season A (i.e., 2015A and 2016A). The second group of environments comprised 12 environments of which seven were from the second growing season, referred to as season B (i.e., 2015B and 2016B). Therefore, each of the two groups of environments was dominated by environments from one of the growing seasons. One outlier environment (Serere2016A) did not fit the classification well (Figure 3.5).

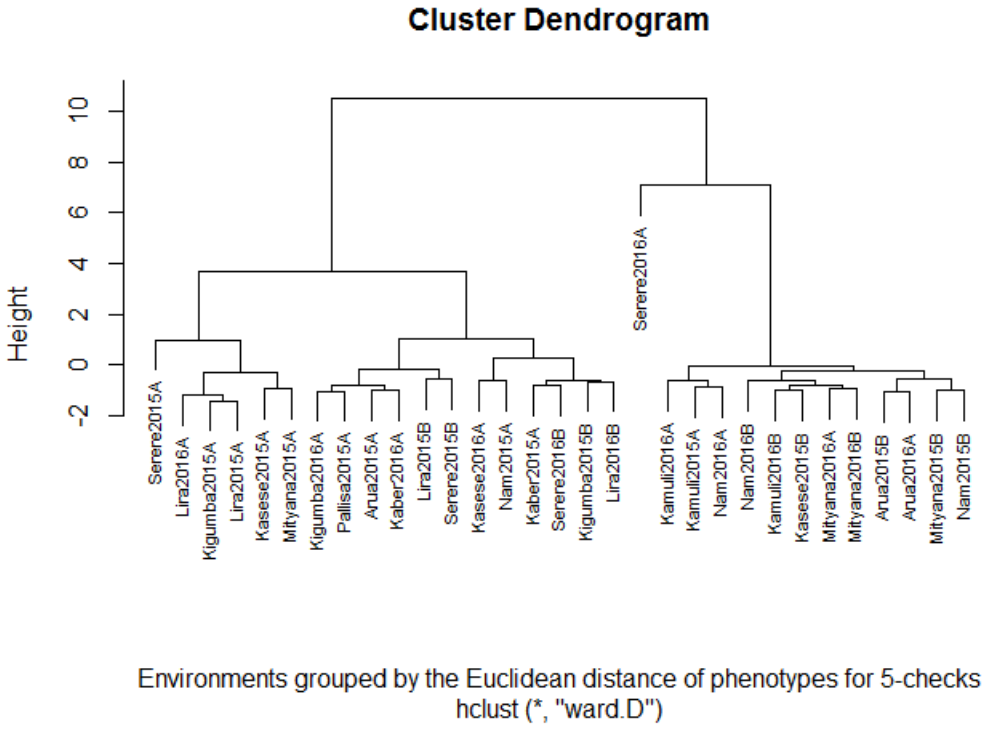


Figure 3.6: Clustering of the environments using nine phenotypic variables for the five-checks, evaluated in all 31 environments (Location-season-year combination)

Three major clusters of environments were observed based on a tree height of 4 from cluster analysis using weather variables (Figure 3.7). Contrary to grouping of environments by check clone phenotypic values, the weather variables grouped the environments by the trial sites. For example, the environments Serere2015B, 2016A, and 2016B were put together into one cluster. Similarly, Mityana2015A, 2015B and 2016A as well as Kigumba 2015A, 2015B and 2016A were grouped into another cluster. Furthermore, correlation of the off-diagonal elements of the distance matrices for phenotypic and weather variable, revealed a non-significant Pearson's correlation coefficient (r = 0.19) between the two variable sets.

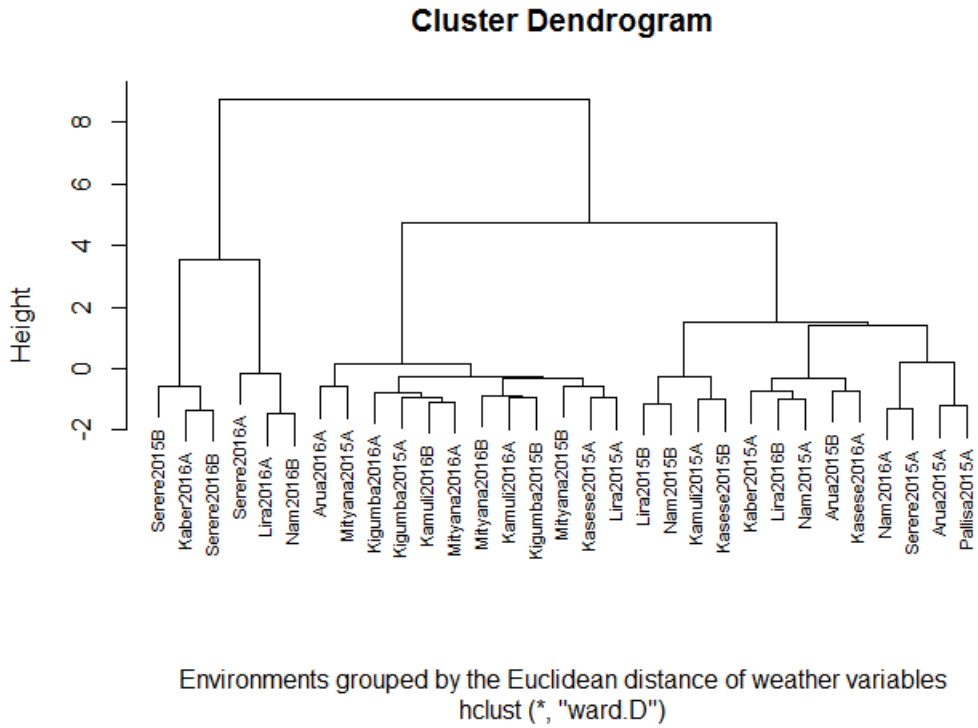


Figure 3.8: Clustering of the environments using the average monthly values computed from the four weather variables (rainfall, temperature, solar radiation and relative humidity)

Testing the relevance of genotype-by-environment interactions

Significant ($P \leq 0.001$) G x E was found for all traits except for CMD (Table 3.1). The deviance values for the full model (genotypic main effects + G x E effects) were lower than for the reduced model (main effects only). Most traits thus had significant G x E, except for CMD scored at three (CMD3s) and six (CMD6s) months after planting (MAP). For CBSD, the

relative proportion of the total phenotypic variance explained by genotypic variance for CBSD6s (25.4%), and CBSDRs (32.1%) were greater than for the G x E and environment main effect variances, except CBSD3s where the variance explained by the environment main effect was 19.8% (Table 3.1). On the other hand, the proportion of the total phenotypic variance explained by the environment-effect was generally higher for RTWT (24.3%), SHWT (29.3), HI (15.2%) and DMC (52.7%) than by the G x E and genotypic main effects, which were less than or equal to 13.1%, except for HI that had 16.0% of the total phenotypic variance attributed to genotypic effects (Table 3.2).

Table 3.3: A chi-square test comparing the deviance values for G x E model (Full Model) with a model fitted without G x E term (Reduced model)

	Deviance values								
Models	CBSD3s	CMD3s	CBSD6s	CMD6s	CBSDRs	SHWT	RTWT	HI	DMC
Full-GxE	6181.0	2246.0	9091.0	3170.5	10479.0	31216.0	28683.0	-4311.0	19816.0
Reduced-NoGxE	6231.0	2244.0	9146.0	3168.5	10610.0	31281.0	28739.0	-4295.0	19854.0
Chi-sq Test	52.2***	0.4^{ns}	56.9***	0.0^{ns}	133.3***	67.7***	58.3***	18.0***	40.1***

*, **, ***, significant at probability levels of 0.05, 0.01, and 0.001 respectively; and ns non-significant.

Table 3.4: Partitioning of the variance components for traits with significant G x E impacts

Traits	Proportion of variance explained by the model predictors (%)		
	Genotype-by-Environment	Environment	Genotype
CBSD3s	12.0	19.8	15.8
CBSD6s	13.5	8.8	25.4
CBSDRs	15.8	1.8	32.1
SHWT	11.9	29.3	8.1
RTWT	9.4	24.3	13.1
HI	7.1	15.2	16.0
DMC	6.8	52.7	6.5

We further assessed the importance of the 48 environmental covariates in explaining the G x E observed for the seven traits showing significant G x E using the variances for genotype reaction norm models (Figure S3.1). Our results did not point out a particular environmental covariate in accounting for the observed G x E across the traits, except for CBSD3s where mean relative humidity (MeanRH01) at the first month of planting explained

reasonable G x E variance (Figure S3.2). Since no specific environmental covariates accounted for the G x E, we used all the 48 environment covariates in the subsequent genomic prediction analyses.

Prediction accuracies for unobserved genotypes across environments

The strategy of predicting newly developed lines using the CV1 scheme showed slightly better predictive ability from the two models (GxECovariates and NoGxECovariates) that explicitly incorporated environmental covariates in the prediction models (Figure 3. 4 and Table S3. 1). The most predictable traits across the three models tested were CBSD3s, CBSD6s and DMC, with accuracies ranging from 0.58 (CBSD6s) for a model with G x E, but no environmental covariates to 0.61 (CBSD6s) for the G x E model that included the environmental covariates. The second most predictable traits were CBSDRs and harvest index with the predictive accuracies that varied from 0.48 for CBSDRs in G x E model, with no covariates (GxENoCovariates) to 0.52 for harvest index in models (GxECovariates and NoGxECovariates) that explicitly incorporated the environmental covariates (Figure 3. 4 and Table S3. 2). The least predictive traits were root and shoot weight per plot (ranging from r = 0.06 for shoot weight to 0.08 for the root weight).

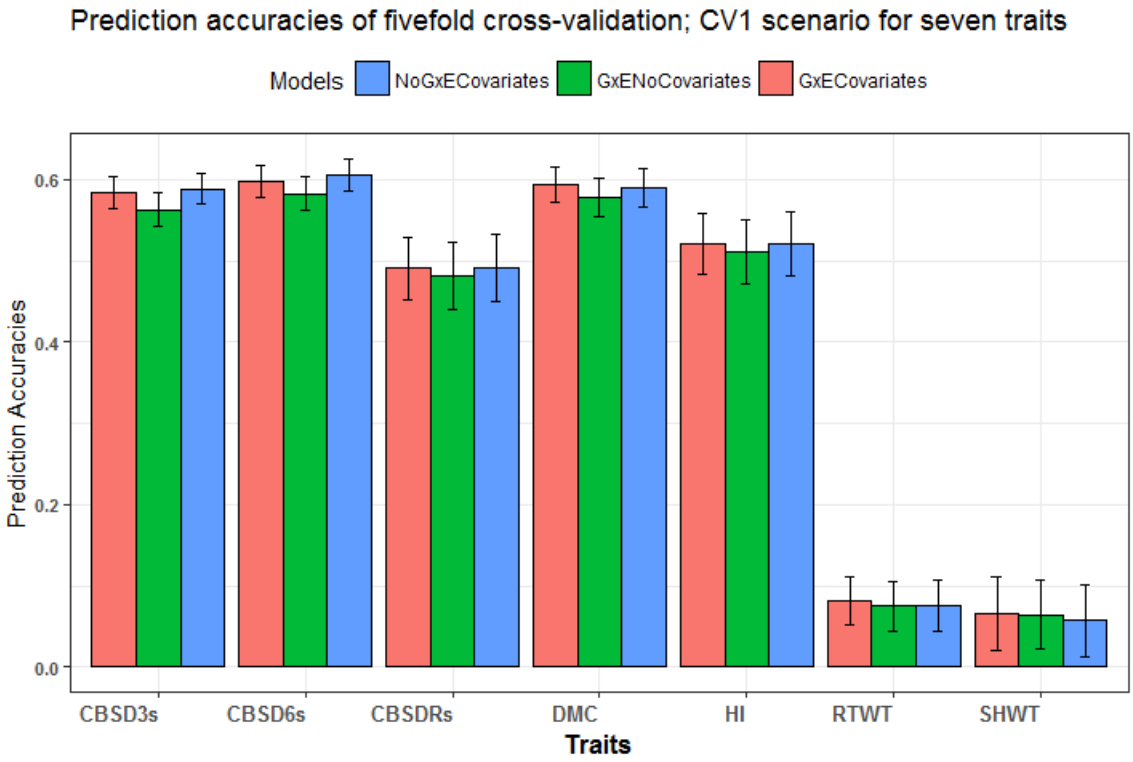


Figure 3.9: The average prediction accuracies for unobserved genotypes across environment. The error bars show the standard deviation of the prediction accuracies for the traits across the three models tested

Prediction accuracies for unobserved genotype in unobserved environment

In general, the model without G x E, but with inclusion of the environmental covariates had higher prediction accuracies than the model that included the interaction term, genomic relationship matrix and variance-covariance matrix of the environmental variables (Figure 3. 5 and Table S3. 2), similar to prediction accuracies observed for unobserved genotypes across environments in CV1 scenario. The most predictable traits were CBSD3s, CBSD6s and DMC, with prediction accuracies ranging from 0.56 for CBSD6s in a G x E model that did not include the environmental covariates (GxENoCovariates) to 0.61 for CBSD6s in a model with environmental covariates, but no G x E term (NoGxECovariates). The prediction accuracies for CBSDRs and HI were moderate, and this varied from 0.45 for CBSDRs in a model with G x E term to 0.52 for HI in a model without G x E. Again, the prediction accuracies were lowest for SHWT and RTWT, ranging from 0.04 for SHWT to 0.08 for RTWT (Table S3. 2).

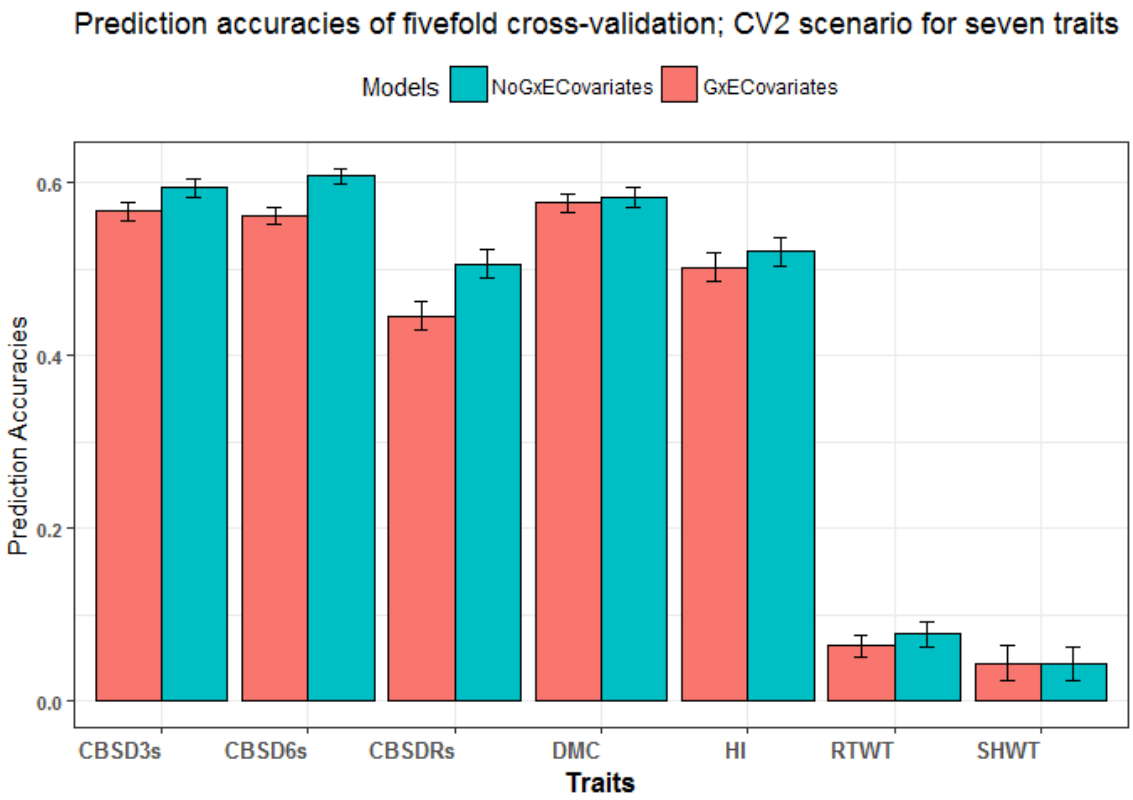


Figure 3.10: Average prediction accuracies from five-fold cross-validation, repeated five times for un observed genotypes in unobserved environments. The error bars show the standard deviation of the prediction accuracies for the traits across the models tested

Prediction accuracies for leave-one-environment-out cross-validation

The results for the prediction analysis where we leave-one-environment-out prediction scheme are presented in Figure 3. 6 and Table S3. 3. For this prediction problem, we only tested the accuracies for the model that included environmental covariates, since the three models did not vary much in prediction accuracies for CV1 and CV2 scenarios. Overall, the prediction accuracies per trait averaged across the 31 environments for CV3 were higher than for the CV1 and CV2 scenarios across traits, ranging from 0.74 for shoot weight per plot to 0.95 for the harvest index. Interestingly, root and shoot weight per plot, which have low heritabilities (Figure 3. 6), observed across location-season-year combination, had high prediction

accuracies of 0.88 and 0.74, respectively (Table S3. 3).

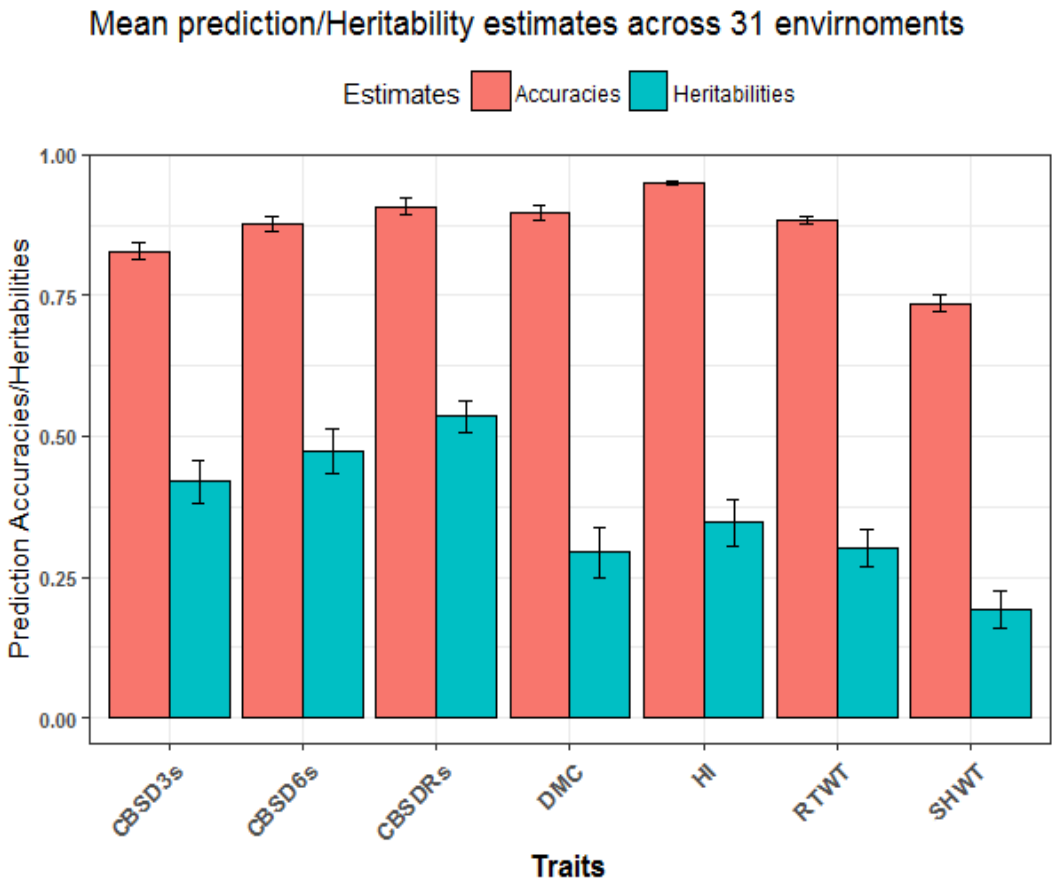


Figure 3. 11: Average prediction accuracies for leave-one-environment-out scenario across 31 environments tested. The error bars show standard deviation for the prediction accuracies across environments for the seven traits.

Discussion

Factors such as variable climate conditions, increasing world population, improvements in diets, and a growing demand for alternative uses for agricultural products are expected to increase the need for more efficient agricultural production (De Leon et al., 2016). In turn, we need to increase the efficiency of breeding to cope with the increasing demand for food production. One approach to improving breeding efficiency is leveraging available genomic resources in combination with environment variables to assess the performance of genotype in their production environments. Genomic selection is one of the tools available in the breeder's toolbox to increase the breeding efficiency of cassava.

Characterization of the environments

According to Bernardo (2003), three approaches are used by the plant breeder to cope with G x E in real breeding situations, one of which is reducing G x E by partitioning the target population environments into smaller more homogenous sub-groups. In this case, the environments in each sub-group might have similar soil type, temperature and precipitation, day length as well as the biotic stresses. Therefore, cultivar recommendations can be made separately for each sub-group of homogenous environments. Based on phenotypic clustering (location-season-year combination), a distinction of the environments by their planting season for the nine traits was observed, with most trials planted in April-May (first season rains) forming a cluster, while the second cluster was dominated by Sep-Oct planting (second season rains). On the contrary, clustering of the environments using weather variables tended to separate the 31 environments by their location. This was evident from weak correlation of 0.19 of the off-diagonal elements of the distance matrices constructed for the two sets of variables (phenotypic and weather). Lack of clear sub-groups based on the two sets of variables, suggest the need for more data collection, especially for edaphic factors that were not measured in the

present study to better delineate the mega-environments. More data generated from major cassava production and consumption zones in Uganda could eventually be fed into a genomic prediction model to increase prediction accuracies.

Genomic prediction accuracy for unobserved genotypes across environments (CV1)

We demonstrated that new experimental lines may be predicted prior to phenotyping using historical data recorded from all sites of interest in the CV1 scheme, which is highly desired by breeders. As genotyping costs continue to decrease, the cost for yield plots either remains constant or increases (Poland and Rife, 2012). Thus, screening hundreds or thousands of seedlings or clones initially through GS becomes a more attractive tool for the breeders. In our study, CV1 showed moderate to high prediction accuracies ranging from 0.48 for CBSDRs in a G x E model where no information sharing was allowed among the environments to 0.61 for CBSD6s in a model with no G x E, but incorporated the environment covariates in the environment main-effect, permitting the sharing of information among environments. Our cross-validation prediction accuracies were higher than previously reported cross-validation prediction accuracies for CBSD3s, CBSD6s and CBSDRs in the NaCRRI training population (Kayondo et al., 2018), suggesting more phenotypic data on genotypes across TPOE and inclusion environmental variates enhanced genomic prediction accuracies. On the other hand, root and shoot weight per plot had poor prediction accuracies across the three models tested. These results suggest GS would be more feasible to screen new genotypes for disease traits, dry matter, and HI as an indirect measure for fresh yield, as proposed by Kawano et al. (1987). We observed only a marginal increase in CV1 prediction accuracies when environmental covariates were incorporated in the genomic prediction models. Despite this, including environmental covariates in GS models is still a reasonable initial step towards understanding the biological relevance of environmental factors leading to the differential response of genotypes across their growing conditions (Bernardo, 2003; Malosetti et al., 2016).

Prediction accuracies for unobserved genotype in unobserved environment (CV2)

This prediction problem is described as the hardest cross-validation scheme, since less information sharing occurs via genotypes or environments (De Leon et al., 2016), yet it provides a practical answer to the plant breeding problem of predicting the performance of never observed genotypes in unevaluated environments. Surprisingly, our cross-validation prediction accuracies for CV2 were similar to the accuracies observed for CV1, with moderate to high prediction accuracies (0.45 to 0.61) observed for CBSD3s, CBSD6s, CBSDRs, DMC and HI and low for RTWT and SHWT. Previously Jarquín et al. (2017) reported much lower cross-validation prediction accuracies for CV2 scheme compared to CV1, in a study predicting the yields of wheat for a Kansas breeding population. Moderate to high prediction accuracies for unobserved genotypes in unobserved environments for CBSD3s, CBSD6s, CBSDRs, HI and DMC further suggests the need to collect and use environmental covariates for genomic prediction, especially for more challenging prediction problem of new genotypes in unobserved environments.

Prediction accuracy of genotypes in the leave-one-environment-out scenario (CV3)

Predicting unobserved environments (CV3) scheme for a G x E model that included covariates produced results that improved the prediction accuracy for all the seven traits (Figure 3. 6 and Table S3. 3). This advantage was greatest for RTWT, where prediction accuracies increased from 0.08 (CV1) and 0.06 (CV2) to 0.88 (CV3). The results were particularly striking for root weight and shoot, given that these traits have low heritabilities (Ozimati et al., 2018; Wolfe et al., 2017). However, for CBSD-related traits and dry matter content with moderate to high heritability (Ozimati et al., 2018), high prediction accuracies observed in this cross-validation scheme (CV3) were not surprising, since most lines had already been observed in many other environments, so we could consider them as replicates of lines in the unobserved environments, thus enabling higher prediction for one missing

environment as the test set. Although this scheme seems the most expensive in terms of the evaluation environments needed for the training set, the level of accuracy observed for RTWT weight in particular, justifies the number of evaluation environments.

Conclusion

This study provides insights into the incorporation of environmental variables into genomic prediction models used in cassava breeding to assess trait performance in light of G x E. Based on the results of the study, CBSD3s and CBSD6s, CBSDRs and HI, DMC are traits for which reasonable prediction accuracies can be achieved in the different prediction problems in cassava breeding such as predicting the performance of newly generated seedlings (crosses) as well as unobserved environments, hence cutting the initial cost of field evaluations.

Acknowledgement

This work was supported by the "Next Generation Cassava Breeding Project" through funds from the Bill and Melinda Gates Foundation and the Department for International Development of the United Kingdom. We thank the technical field staff of NaCRRI Cassava-breeding programme, who helped in the trial management and/or data collection (Joseph Orone, Charles Majara, Gerald Adiga,Vincent Kyaligonza). Thank you to Andrew Ssali, the meteorologist attached to NaCRRI, who helped in installation and retrieving of the weather data. We thank the laboratory staff, particularly Francis Osingada and Jimmy Akano, for the support they provided during DNA extraction and quantification. Lastly, we thank Cornell University Institute of Genomic diversity (IGD) for carrying out the genotyping.

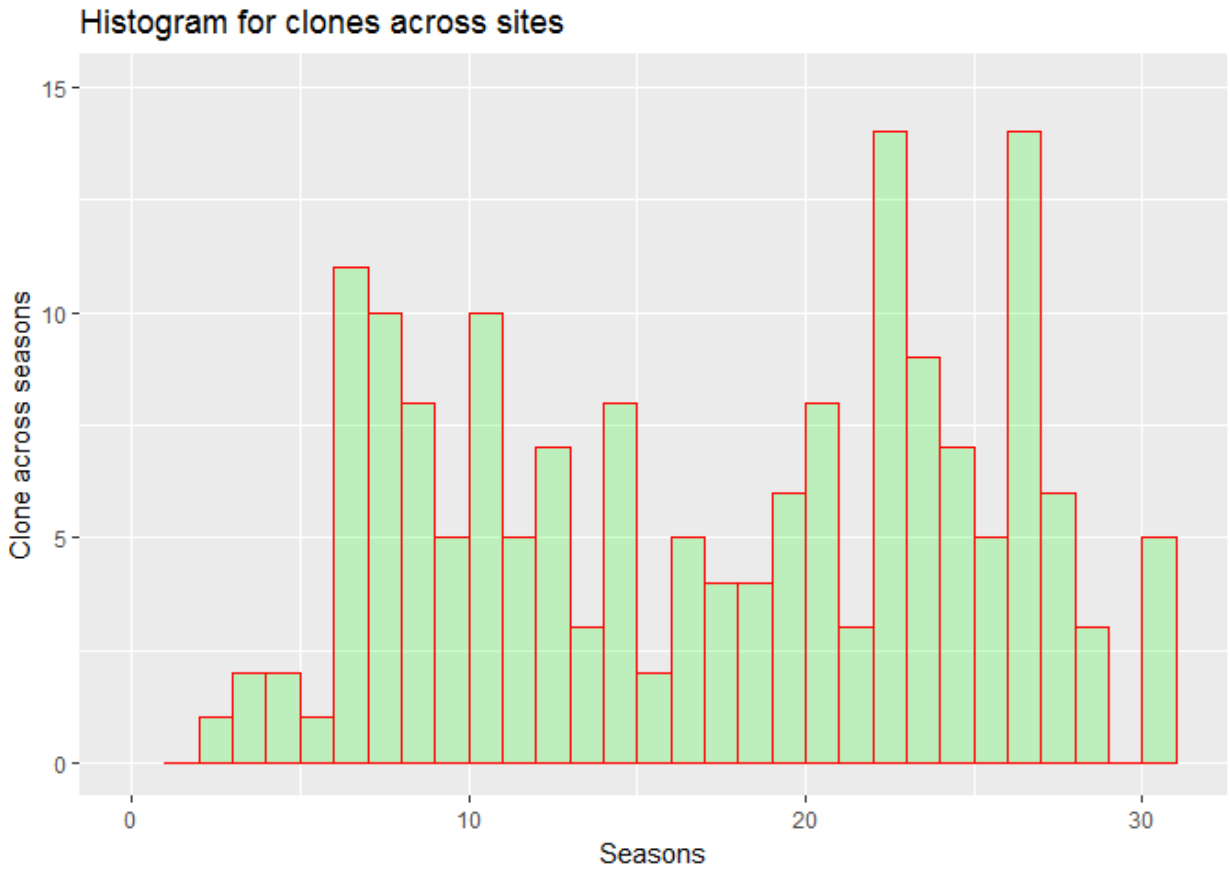
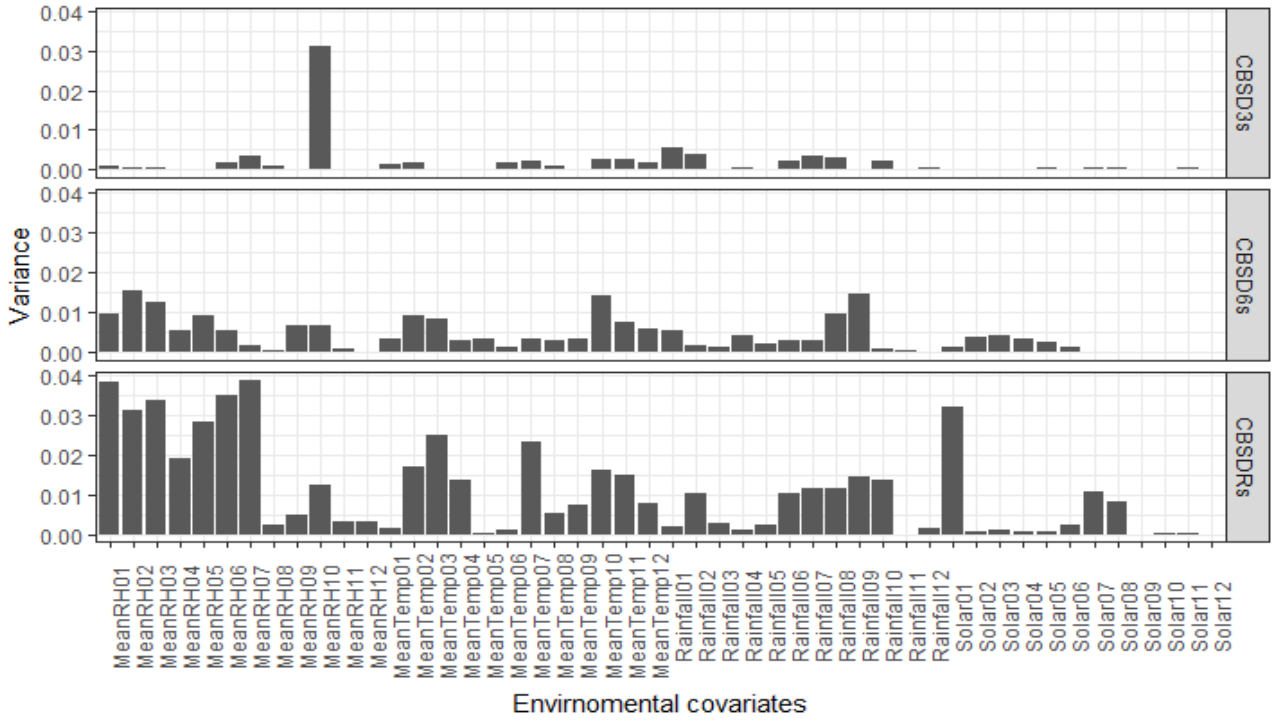


Figure S3.3: The number of times the clone appears across the 31 environments evaluated. Only the five checks were evaluated in all 31 environments



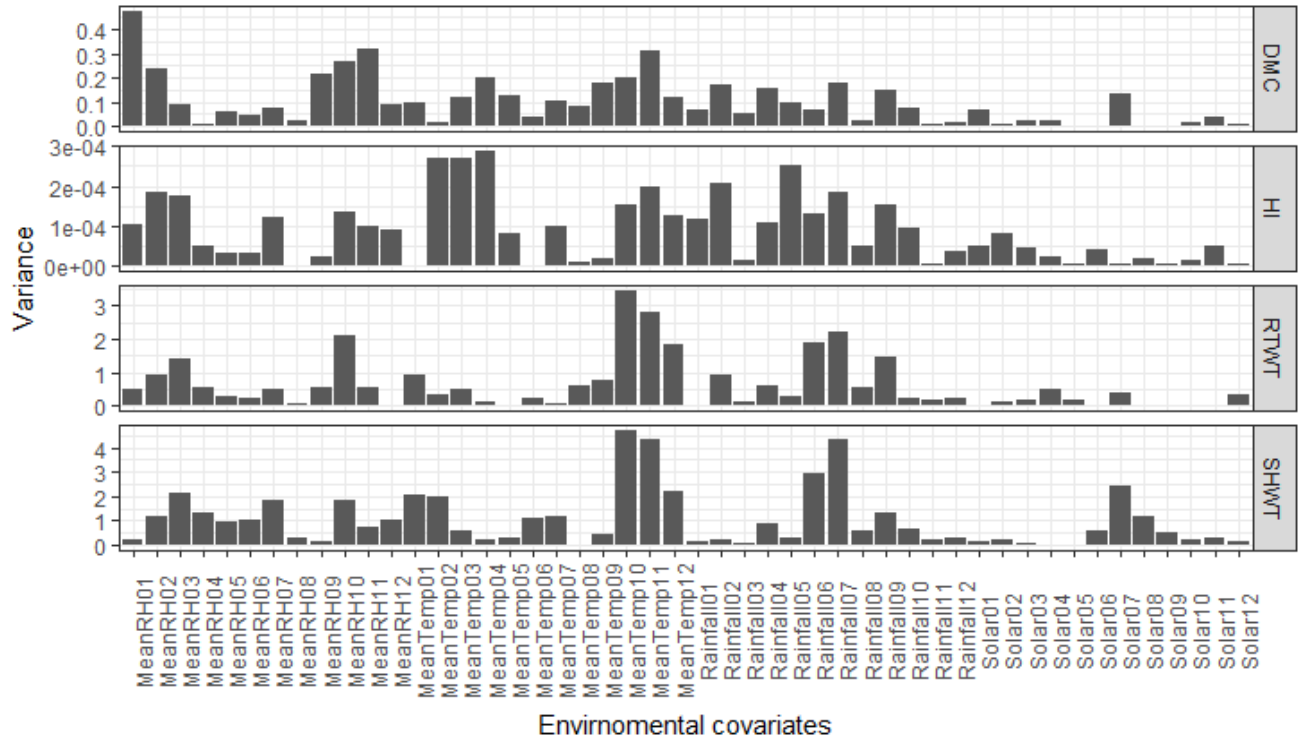


Figure S3. 4: Variance of the reaction norm for the genotypes explained by each weather variables as a covariate to assess their relative importance in accounting for G x E observed for the 7 traits

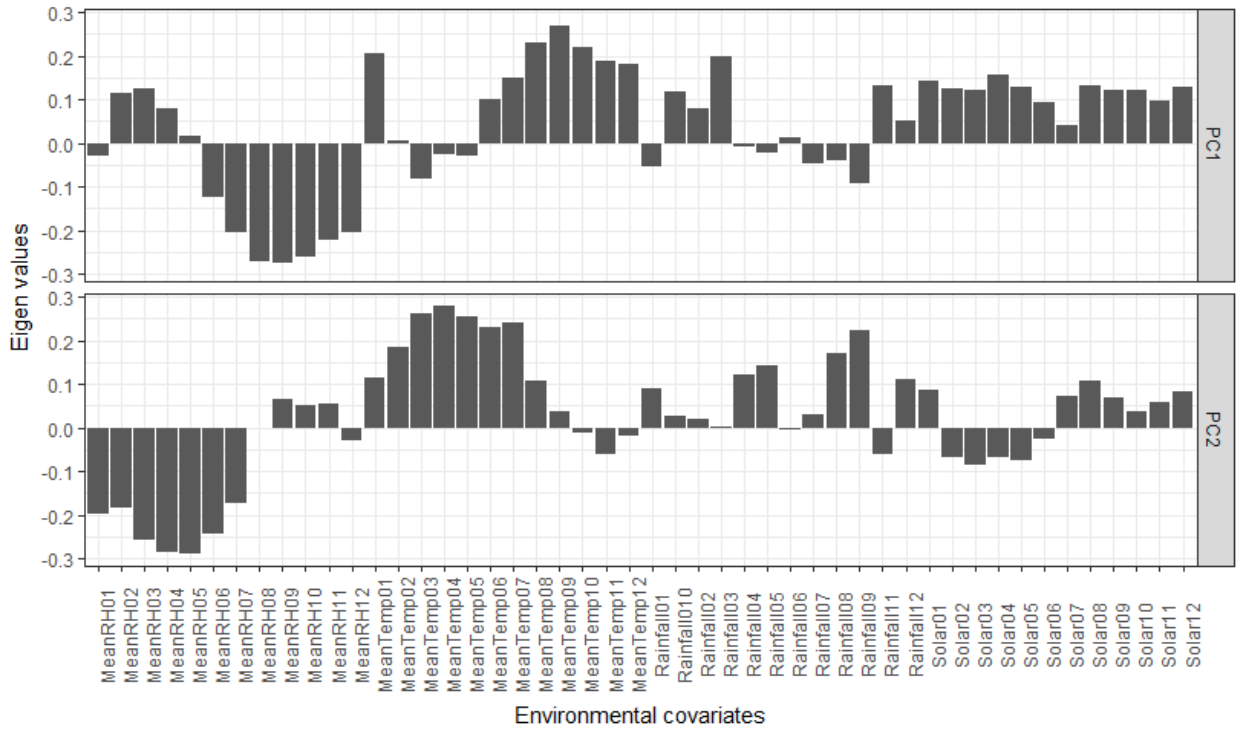


Figure S3.5: PC1 and PC2 loadings of the 48 environmental variables across 31 environments

Table S3. 3: Mean prediction accuracies for CV1 scenario of predicting newly developed genotypes or introduced germplasm.

Traits	Models	Mean Accuracies	S.D	S.E	C.I
CBSD3s	GxECovariates	0.58	0.10	0.02	0.04
CBSD3s	GxENoCovariates	0.56	0.10	0.02	0.04
CBSD3s	NoGxECovariates	0.59	0.10	0.02	0.04
CBSD6s	GxECovariates	0.61	0.10	0.02	0.04
CBSD6s	GxENoCovariates	0.58	0.11	0.02	0.04
CBSD6s	NoGxECovariates	0.60	0.10	0.02	0.04
CBSDRs	GxECovariates	0.49	0.19	0.04	0.08
CBSDRs	GxENoCovariates	0.48	0.20	0.04	0.08
CBSDRs	NoGxECovariates	0.49	0.21	0.04	0.09
DMC	GxECovariates	0.59	0.11	0.02	0.05
DMC	GxENoCovariates	0.58	0.12	0.02	0.05
DMC	NoGxECovariates	0.59	0.12	0.02	0.05
HI	GxECovariates	0.52	0.19	0.04	0.08
HI	GxENoCovariates	0.51	0.20	0.04	0.08
HI	NoGxECovariates	0.52	0.20	0.04	0.08
RTWT	GxECovariates	0.08	0.15	0.03	0.06
RTWT	GxENoCovariates	0.07	0.15	0.03	0.06
RTWT	NoGxECovariates	0.08	0.16	0.03	0.07
SHWT	GxECovariates	0.07	0.23	0.05	0.09
SHWT	GxENoCovariates	0.06	0.21	0.04	0.09
SHWT	NoGxECovariates	0.06	0.22	0.04	0.09

Table S3. 4: Mean prediction accuracies for five-fold and five repeats cross-validation strategy for unobserved genotype in unobserved environments (CV2).

Traits	Model	Mean Accuracies	S.D	S.E	C.I
CBSD3s	GxECovariates	0.57	0.14	0.01	0.02
CBSD3s	NoGxECovariates	0.59	0.13	0.01	0.02
CBSD6s	GxECovariates	0.56	0.12	0.01	0.02
CBSD6s	NoGxECovariates	0.61	0.10	0.01	0.02
CBSDRs	GxECovariates	0.45	0.20	0.02	0.03
CBSDRs	NoGxECovariates	0.51	0.21	0.02	0.03
DMC	GxECovariates	0.58	0.14	0.01	0.02
DMC	NoGxECovariates	0.58	0.14	0.01	0.02
HI	GxECovariates	0.50	0.20	0.02	0.03
HI	NoGxECovariates	0.52	0.20	0.02	0.03
RTWT	GxECovariates	0.06	0.16	0.01	0.03
RTWT	NoGxECovariates	0.08	0.18	0.01	0.03
SHWT	GxECovariates	0.04	0.25	0.02	0.04
SHWT	NoGxECovariates	0.04	0.24	0.02	0.04

Table S3. 5: Prediction accuracies for leave-one-environment-out scenario (CV3)

Traits	Mean Accuracies	S.D	SE	C.I
CBSD3s	0.83	0.09	0.02	0.03
CBSD6s	0.88	0.08	0.01	0.03
CBSDRs	0.91	0.08	0.02	0.03
DMC	0.90	0.03	0.01	0.03
HI	0.95	0.02	0.00	0.01
RTWT	0.88	0.04	0.01	0.02
SHWT	0.74	0.08	0.02	0.03

References

Akdemir, D., and U.G. Okeke. 2015. EMMREML: Fitting mixed models with known covariance structures, R Repository CRAN.

Bernardo, R. 2003. Breeding for Quantitative Traits in Plants. Stemma Press, Woodbury, Minnesota.

Burgueño, J., G. de los Campos, K. Weigel, and J. Crossa. 2012. Genomic prediction of breeding values when modeling genotype x environment interaction using pedigree and dense molecular markers. Crop Sci. 52(2): 707–719.

Ceballos, H., J.C. Pérez, F. Calle, G. Jaramillo, J.I. Lenis, N. Morante, and J. López. 2007. A new evaluation scheme for cassava breeding at CIAT. : 125–135.

Cooper, M., C.D. Messina, D. Podlich, L.R. Totir, A. Baumgarten, N.J. Hausmann, D. Wright, and G. Graham. 2014. Predicting the future of plant breeding: Complementing empirical evaluation with genetic prediction. Crop Pasture Sci. 65(4): 311–336

Crossa, J., P. Perez, G. Mahuku, C. Magorokosho, I. Maize, and C. De Postgraduados. 2010. Genomic Prediction of Quantitative Traits in Plant. Cosmos (Mm): 1–33.

Elshire, R.J., J.C. Glaubitz, Q. Sun, J. a Poland, K. Kawamoto, E.S. Buckler, and S.E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6: 5–13.

Endelman, J.B. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome J. 4: 250.

Esuma, W., R.S. Kawuki, L. Herselman, and M.T. Labuschagne. 2016. Stability and

genotype by environment interaction of provitamin A carotenoid and dry matter content in cassava in Uganda. Breed. Sci. 66(3): 434–443.

Federer, W.T and N. K. Nguyen. 1954: Constructing Augmented Experiment Designs with Gendex. *Biometrics Unit Tech. Reports* **BU-1610-M**.

Hamblin, M.T., and I.Y. Rabbi. 2014. The effects of restriction-enzyme choice on properties of genotyping-by-sequencing libraries: A study in Cassava (*Manihot esculenta*). Crop Sci. 54: 2603–2608.

Hayes, B.J., P.J. Bowman, A.J. Chamberlain, and M.E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: progress and challenges. J. Dairy Sci. 92: 433–443.

Heslot, N., D. Akdemir, M.E. Sorrells, and J.L. Jannink. 2014. Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theor. Appl. Genet. 127(2): 463–480.

Hillocks, R.J., and J.M. Thresh. 2000. Cassava Mosaic and Cassava Brown Streak Virus Diseases in Africa : Root 7: 1–8.

IITA. 1990. Cassava in Tropical Africa: A reference manual. 3: 1–176.

Jarquín, D., J. Crossa, X. Lacaze, P. Du Cheyron, J. Daucourt, J. Lorgeou, F. Piraux, L. Guerreiro, P. Pérez, M. Calus, J. Burgueño, and G. de los Campos. 2014. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor. Appl. Genet. 127(3): 595–607.

Jarquín, D., C. Lemes da Silva, R.C. Gaynor, J. Poland, A. Fritz, R. Howard, S. Battenfield, and J. Crossa. 2017. Increasing Genomic-Enabled Prediction Accuracy by Modeling Genotype × Environment Interactions in Kansas Wheat. Plant Genome 10(2)

Kawano, K., W.M.G. Fukuda, and U. Cenpukdee. 1987. Genetic and Environmental Effects on Dry Matter Content of Cassava Root. Crop Sci. 27: 69.

Kayondo, S.I., D.P. Del Carpio, R. Lozano, A. Ozimati, M. Wolfe, Y. Baguma, V. Gracen, S. Offei, M. Ferguson, R. Kawuki, and J.L. Jannink. 2018. Genome-wide association mapping and genomic prediction for CBSD resistance in Manihot esculenta. Sci. Rep. 8: 1–11.

Lado, B., P.G. Barrios, M. Quincke, P. Silva, and L. Gutirrez. 2016. Modeling genotype- by-environment interaction for genomic selection with unbalanced data from a wheat breeding program. Crop Sci. 56(5): 2165–2179.

De Leon, N., J.-L. Jannink, J.W. Edwards, and S.M. Kaeppler. 2016. Introduction to a

Special Issue on Genotype by Environment Interaction. Crop Sci. 56(5): 2081.

De Los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, and J.M. Cotes. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182: 375–385.

Ly, D., M. Hamblin, I. Rabbi, G. Melaku, M. Bakare, H.G. Gauch, R. Okechukwu, A.G.O. Dixon, P. Kulakow, and J.L. Jannink. 2013. Relatedness and genotype-by-environment interaction affect prediction accuracies in genomic selection: A study in cassava. Crop Sci. 53(4): 1312–1325.

Ly, D., S. Huet, A. Gauffreteau, R. Rincent, G. Touzy, A. Mini, J.L. Jannink, F. Cormier, E. Paux, S. Lafarge, J. Le Gouis, and G. Charmet. 2018. Whole-genome prediction of reaction norms to environmental stress in bread wheat (*Triticum aestivum* L.) by genomic random regression. F. Crop. Res. 216(December 2016): 32–41.

Lynch, M., and B. Walsh. 1998. Genetics and Analysis of Quantitative Traits. Genet. Anal. Quant. Trait. 2: 980.

Malosetti, M., D. Bustos-Korts, M.P. Boer, and F.A. Van Eeuwijk. 2016. Predicting responses in multiple environments: Issues in relation to genotype-by-Environment interactions. Crop Sci. 56(5): 2210–2222.

Malosetti, M., J.M. Ribaut, and F.A. van Eeuwijk. 2013. The statistical analysis of multi-environment data: Modeling genotype-by-environment interaction and its genetic basis. Front. Physiol. 4 MAR(March): 1–17.

Masinde, E.A., G. Mkamillo, J.O. Ogendo, R. Hillocks, R.M.S. Mulwa, B. Kimata, and M.N. Maruthi. 2018. Genotype by environment interactions in identifying cassava (*Manihot esculenta* Crantz) resistant to cassava brown streak disease. F. Crop. Res. 215(September 2017): 39–48.

Meuwissen, T. H. E. , Hayes, B. J., & Goddard, M.E. 2001. Prediction of total genetic value using genome-wide dense markers maps. Genetics 157: 1819–1829.

Mtunguja, M.K., H.S. Laswai, E. Kanju, J. Ndunguru, and Y.C. Muzanila. 2016. Effect of genotype and genotype by environment interaction on total cyanide content, fresh root, and starch yield in farmer-preferred cassava landraces in Tanzania. Food Sci. Nutr. 4(6): 791–801.

Murtagh, F., and P. Legendre. 2011. Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm. June: 1–20.

Okeke, U.G., D. Akdemir, I. Rabbi, P. Kulakow, and J.L. Jannink. 2017. Accuracies of

univariate and multivariate genomic prediction models in African cassava. Genet. Sel. Evol. 49(1): 1–10.

Ozimati, A., R. Kawuki, W. Esuma, S. I. Kayondo, A. Pariyo, M. Wolfe, and J-L. Jannink. 2018. Genetic Variation and Trait Correlations in East African Cassava Breeding Population for Genomic Selection in review

Pariyo, A., Y. Baguma, T. Alicai, R. Kawuki, E. Kanju, A. Bua, C.A. Omongo, P. Gibson, D.S. Osiru, and D. Mpairwe. 2015. Stability of resistance to cassava brown streak disease in major agro-ecologies of Uganda. J. Plant Breed. Crop Sci. 7: 67–78.

Poland, J.A., and T.W. Rife. 2012. Genotyping-by-Sequencing for Plant Breeding and Genetics. Plant Genome J. 5(3): 92.

Prochnik, S., P.R. Marri, B. Desany, P.D. Rabinowicz, C. Kodira, M. Mohiuddin, F. Rodriguez, C. Fauquet, J. Tohme, T. Harkins, D.S. Rokhsar, and S. Rounsley. 2012. The Cassava Genome: Current Progress, Future Directions. Trop. Plant Biol. 5: 88–94.

R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. R Found. Stat. Comput. Vienna, Austria (ISBN 3-900051-07-0): 900051.

Tumuhimbise, R., R. Melis, P. Shanahan, and R. Kawuki. 2014. Genotype x environment interaction effects on early fresh storage root yield and related traits in cassava. Crop J. 2(5): 329–337.

Van Eeuwijk, F.A., D. V. Bustos-Korts, and M. Malosetti. 2016. What should students in plant breeding know about the statistical aspects of genotype-by-Environment interactions Crop Sci. 56(5): 2119–2140.

Wolfe, M.D., I.Y. Rabbi, C. Egesi, M. Hamblin, R. Kawuki, P. Kulakow, R. Lozano, D.P. Del Carpio, P. Ramu, and J.-L. Jannink. 2016. Genome-Wide Association and Prediction Reveals Genetic Architecture of Cassava Mosaic Disease Resistance and Prospects for Rapid Genetic Improvement. Plant Genome 9: 342–356.

Yan, W., M.S. Kang, B. Ma, S. Woods, and P.L. Cornelius. 2007. GGE biplot vs. AMMI analysis of genotype-by-environment data. Crop Sci. 47(2): 643–655.

CHAPTER 4

TRAINING POPULATION OPTIMIZATION FOR PREDICTION OF CASSAVA BROWN STREAK DISEASE RESISTANCE IN WEST AFRICAN CLONES

Abstract

Cassava production in the central, southern and eastern parts of Africa is under threat by cassava brown streak virus (CBSV). Yield losses of up to 100% occur in cases of severe infections of edible roots. Easy illegal movement of planting materials across African countries, and long-range movement of the virus vector (*Bemisia tabaci*) may facilitate spread of CBSV to West Africa. Thus, effort to pre-emptively breed for CBSD resistance in W. Africa is critical. Genomic selection (GS) has become the main approach for cassava breeding, as costs of genotyping per sample have declined. Using phenotypic and genotypic data (genotyping-by-sequencing), followed by imputation to whole genome sequence (WGS) for 922 clones from National Crops Resources Research Institute, Namulonge, Uganda as a training population (TP), we predicted CBSD symptoms for 35 genotyped W. African clones, evaluated in Uganda. The highest prediction accuracy (r = 0.44) was observed for cassava brown streak disease severity scored at three months (CBSD3s) in the W. African clones using WGS-imputed markers. Optimized TPs gave higher prediction accuracies for CBSD3s and CBSD6s than random TPs of the same size. Inclusion of CBSD QTL chromosome markers as kernels, increased prediction accuracies for CBSD3s and CBSD6s. Similarly, WGS imputation of markers increased prediction accuracies for CBSD3s and for cassava brown streak disease root severity (CBSDRs), but not for CBSD6s. Based on these results we recommend TP optimization, inclusion of CBSD QTL markers in genomic prediction models, and the use of high-density (WGS-imputed) markers for CBSD predictions across population.

Manuscript accepted for publication in G3: Alfred Ozimati, Robert Kawuki, Williams Esuma, Ismail Siraj Kayondo, Marnin Wolfe, Roberto Lozano, Ismail Rabbi, Peter Kulakow, and Jean-Luc Jannink. 2018. Training Population Optimization for Prediction of Cassava Brown Streak Disease Resistance in West African Clones.

Introduction

Cassava (*Manihot esculenta* Crantz) is ranked the fourth most important source of calories in the developing world, after wheat, maize, and rice, and is estimated to feed a population of about 700 million people directly or indirectly (Legg et al., 2014). Reports on global cassava production in the 1960's positioned Brazil as the leading producer in the world, however in the 1990's Nigeria became the world's largest cassava producer, accounting for half of the world's total production (Nweke, 2004). Other African countries where cassava is a major staple food crop include Uganda, Tanzania and Kenya in eastern Africa, Malawi and Mozambique in southern Africa, Democratic Republic of Congo (DRC) in central Africa, and Ghana in western Africa (Hillocks and Jennings, 2003). Cassava is popular in Africa as a food security crop, because of its resilience under drought and poor soils, and its ability to be easily propagated through stem cuttings (Masona et al., 2001; Legg et al., 2014).

Yields of cassava have remained low (8-12 tons/ha) in Africa compared to Asian countries such as Thailand and Vietnam where yield averaged are up to 20 tons/ha (Nweke, 2004). Reasons for relatively low yields in Africa include both abiotic (low soil fertility and socio-economic factors such as lack of access to improved varieties) and biotic factors (Nweke, 2004). The most devastating biotic stresses today are the cassava brown streak (CBSD) and cassava mosaic (CMD) diseases (Maruthi et al., 2005; Mware et al., 2009). Of these two virus-induced diseases, CBSD is the most important constraint to cassava production in central, eastern and southern Africa as it causes yield losses of up to 100% (Alicai et al., 2007; Hillocks et al., 2016).

Phylogenetic analysis of complete viral RNA genome sequences taken from CBSD symptomatic plants, sampled across eastern and southern Africa, revealed two clades of distinct CBSD-causing virus species that were named: Uganda cassava brown streak virus (UCBSV) and cassava brown streak virus (CBSV) (Winter et al., 2010; Mohammed et al., 2011; Patil et

al., 2015; Alicai et al., 2016; Mbewe et al., 2017). The two species belong to genus *Ipomovirus* within the family of *Potyviridae*, and share an identity of 70% and 74% at the level of nucleotide and polyprotein amino acid sequences, respectively (Monger et al., 2001; Winter et al., 2010). Cassava brown streak disease symptoms on cassava leaves manifest as feathery chlorosis around secondary veins, which may disappear when new growth starts after a period of drought-induced leaf abscission (Hillocks, 2004). While on the roots, CBSD symptoms externally present as radial constriction, and internally as brown necrotic lesions on part or all of the starchy root, making it inedible (Hillocks, 2004; Hillocks et al., 2016).

Although the first incidence of CBSD was reported in 1930's (Storey and Nichols, 1938), little attention was paid to it, because geographically CBSD was confined to the low altitudes of east African coastal region (less than 1000 m.a.s.l). Nonetheless, CBSD has spread rapidly to other countries including; Uganda, Burundi, DRC, Mozambique and Rwanda in the last 2 decades to cover wider range of altitudes than previously reported (Hillocks et al., 2002; Alicai et al., 2007; Legg et al., 2011; Mulimbi et al., 2012). Cassava brown streak disease is commonly spread through sharing of infected stem cuttings for propagation, in addition to super-abundant whitefly *Bemisia tabaci*, as a vector (Hillocks and Jennings, 2003; Njoroge et al., 2017).

Officially, genetic materials can move from W. Africa to E. Africa, but movement in the reverse direction is prohibited to prevent accidental introduction of CBSD-causing viruses in W. Africa. Nevertheless, the free movement of planting materials across farming communities has led to increased fear that CBSD could spread to other regions, including West Africa (Legg et al., 2014; Patil et al., 2015; Beyene et al., 2017). Given the current impact of CBSD on cassava production in endemic countries, effort needs to be in place to avert or minimize future CBSD impact in W. Africa, especially Nigeria the world's leading cassava

producer. Among other methods, Legg et al., (2014) proposed pre-emptive breeding for CBSD resistant clones in W. Africa.

High levels of field resistance to CBSD have been reported from genetically transformed plants with coat protein of UCBSV and CBSV, compared to non-transformed plants (Ogwok et al., 2012; Odipio et al., 2014; Beyene et al., 2017; Wagaba et al., 2017). However, the transgenic CBSD resistant clones are still within research confinement, because of unclear regulatory frameworks regarding field production of genetically modified organisms (GMO) in Uganda and east Africa at large. Other efforts to breed for CBSD resistance in E. Africa are geared towards identification of quantitative trait loci (QTL) for CBSD resistance, with the aim of developing molecular markers to implement marker assisted selection (MAS). A number of QTL mapping studies for CBSD resistance in E. African germplasm have been conducted, and the studies pointed out both unique and overlapping QTL regions for which markers could be developed for MAS (Kayondo et al., 2018; Masumba et al., 2017; Nzuki et al., 2017). One of the highest effect QTL detected involved a bi-parental mapping population from a cross between Kiroba and AR37-80 that explained 18% of total phenotypic variance (Nzuki et al., 2017). However, using bi-parental QTL to develop markers for MAS is only feasible if the QTL are validated in other breeding populations. Furthermore, the recent genome-wide association studies conducted by Kayondo et al. (2018), using same training populations (TP1 and TP2), confirmed the polygenic nature of CBSD resistance previously reported (Kawuki et al., 2016).

Genomic selection, proposed by Meuwissen et al. (2001) provides an option for using DNA markers for traits that are truly quantitative, where no single causal locus accounts for a major fraction of the variation for selection decisions. Genomic selection (GS) relies on a genome-wide distribution of markers to ensure all QTL have at least one marker in high LD, enabling selection on highly polygenic traits. Genomic selection is typically done using a

phenotyped and genotyped training population to estimate genome-wide marker effects (Hayes et al., 2009). The genomic estimated breeding values (GEBV) for all genotyped individuals can then be computed as the sum of marker effects multiplied by the marker genotypes across the whole genome (Meuwissen, et al., 2001). These GEBVs aim to capture all QTL accounting for variation in target traits (Hayes et al., 2009).

Although GS has reportedly outperformed traditional selection methods such MAS and marker assisted recurrent selection (MARS) for quantitative traits (Goiffon, 2016), successful implementation of genomic selection depends on a number of factors including: trait heritability, marker density, the size of the training population, the relationship between the training population (TP) and the selection candidates (Jannink et al., 2010; Heffner et al., 2011; Nielsen et al., 2016).

Increases in prediction accuracy have been reported by composing training populations from optimal subsets of individuals chosen to minimize the expected prediction error variance (PEV) of the selection candidates compared to using random subsets or even the full set of available individuals (Rincent et al., 2012; Akdemir et al., 2015; Isidro et al., 2015; Yu et al., 2016). Furthermore, studies have shown increased prediction accuracies with inclusion of prior QTL information in genomic prediction models. For example, a study by Haffliger, (2016) for reproductive traits in Swiss pig breeds revealed a significant increase in prediction accuracy for piglets when previously detected reproductive trait QTL markers were included in the prediction model.

Thus, this study aimed to evaluate the use of genomic predictions of West African clones using training data from a Ugandan population as a pre-emptive breeding strategy for CBSD resistance. Specifically we tested CBSD prediction accuracies for (i) different sizes of training populations across genomic prediction models (ii) random and optimized training sets,

(iii) models with incorporation of prior CBSD QTL, and (iv) high and low density marker panels.

Materials and Methods

Constitution and evaluation of training population

The training population comprised 922 clones, combined from two experimental trials. For consistency, we refer to the trials as training population 1 (TP1) and training population 2 (TP2). A total of 400 clones constituting TP1 were generated from crossing diverse parents that were assembled from International Center for Tropical Agriculture (CIAT), International Institute of Tropical Agriculture (IITA), Tanzania, and National Crops Resources Research Institute (NaCRRI), Uganda. The introductions from CIAT targeted improvement for quality and yield traits, while the germplasm from the IITA, Tanzania and NaCRRI breeding programs targeted resistance to CBSD.

Crosses were made among the progenitors to generate TP1 in 2009-2010, from which both controlled crosses and open-pollinated seeds were harvested. After seedling evaluation, the first clonal evaluation for TP1 was done at Namulonge in 2012-2013 by conducting an un-replicated experiment, and afterwards expanded to 3 sites (Kasese, Ngetta and Namuloge) for the second year of clonal evaluation, planted in alpha lattice design, in single row plots of 10 plants, replicated twice.

Meanwhile, the second set of training population (TP2) comprised 522 clones, generated from open-pollinated seeds that were harvested from the first clonal evaluation trials of TP1. Similar to TP1, after a year of seedling evaluation (2013-2014), TP2 was planted for the first clonal evaluations in 2014 at 2 sites (Namulonge and Kamuli). In 2015, TP2 was replanted for the second year of clonal evaluation, with the trials expanded to 3 sites (Namulonge, Kamuli, Serere). Thus, Namulonge was the only overlapping evaluation site between TP1 and TP2. The clonal evaluations for TP2 were established in an augmented incomplete block design with six common checks per block, and each plot within a block

containing 10 plants established in a single row. Planting of all the trials was done at spacing of 1 m x 1 m adopted within and between rows, while blocks were separated by 2 m alleys.

Data on foliar CBSD severity was collected at three and six months after planting (MAP), while the roots were evaluated for CBSD severity at 12 MAP. Foliar severity for CBSD was assessed on a scale of 1-5 (Hillocks and Thresh, 2000), where: 1 = no symptom; 2 = slight foliar chlorotic leaf mottle with no stem lesions; 3 = foliar chlorotic leaf mottle and blotches with mild stem lesions, but no die back; 4 = foliar chlorotic leaf mottle and blotches with pronounced stem lesions, but no die back; and 5 = defoliation with stem lesions and dieback. To assess root necrosis severity, each root was sliced transversely 5-7 times and the cross-sections scored for necrotic symptoms on a scale of 1-5 (Hillocks and Thresh, 2000), where: 1 = no necrosis, 2 = ≤ 5% necrotic; 3 = 6-10% necrotic; 4 = 11-25% necrotic and mild root constriction; and 5 = >25% necrotic and severe root constriction.

West African genetic materials and evaluation

In 2015, we received a total of 95 clones that constituted part of IITA, Nigeria genetic gain population for implementing genomic selection (Wolfe et al., 2017). These clones were shipped to Uganda in the form of tissue culture plantlets. The first set of 30 clones was received in February 2015 and the second lot of 65 clones was received in June 2015. The plantlets were multiplied in tissue culture and further hardened in a screen house for three months.

In August and November 2015, the first set of 30 and the second set of 65 clones were planted in the field at Namulonge. This was done to further generate adequate stem cuttings for establishment of standard experiment. In September 2016, after generating enough planting material, we established a trial for the first set of 27 clones that survived, in a randomized complete block design (RCBD) replicated twice, with each plot containing 10 plants in a single row.

For the second set of 65 clones, unfortunately we lost more than half of the clones due to drought that occurred a month after their first field exposure in 2015. The remaining 22 clones that survived were planted in November 2016, again using an RCBD, replicated twice. In contrast to the first set of 27 clones, there was only enough planting material for 5 plants per plot. Cassava brown streak disease phenotyping was taken as described previously, for the two TPs and all infections occurred under natural conditions.

DNA extraction and genotyping

Approximately 100 mg of fresh tissue was collected from tender apical leaves of TP1 and TP2 clones for DNA extraction. DNA was extracted following the protocol for the QIAGEN DNeasy extraction kit and quantified using the PicoGreen® DNA quantification kit to ensure the required concentrations were obtained for sequencing. The extracted DNA samples were shipped to the Cornell University Genomics Diversity Facility for genotyping, using the genotyping-by-sequencing (GBS) approach (Elshire et al., 2011). The GBS libraries were constructed using the ApeKI restriction enzyme as described previously (Rabbi et al., 2014).

Marker genotypes were called using TASSEL GBS pipeline v4 (Glaubitz et al., 2014), after aligning the reads to the Cassava reference genome v6 (Prochnik et al., 2012). Using VCFtools, Variant Calling Format (VCF) files were generated for each chromosome. Genotypes with less than five reads were masked before imputation. Similarly, markers with more than 60% missing calls were removed. Only bi-allelic GBS SNP markers were considered for further processing. Missing markers were imputed using Beagle 4.1 software (Browning and Browning, 2016), with default parameter settings. In all, 46,760 SNPs remained, which we referred to as the "GBS" markers.

In addition, we used a second set of markers, which we referred to as "whole genome sequence" (or "WGS") imputed markers. The WGS markers were imputed using IMPUTE-2

software described in detail by Lozano et al., (2017). Briefly, the WGS imputation relied on the cassava HapMapII, a collection of 241 genome-sequenced samples (Ramu et al., 2017) as a reference panel. This reference panel comprised mainly of improved cassava clones under cultivation and a few wild relatives, and contained 28 million SNP markers (Lozano et al., 2017). The set of markers referred to as "WGS" (or "high density") hereafter, included ~5 million SNPs.

Statistical analyses

Analyses of phenotypic data

Because of differences in trial design for TP1 and TP2 as well as the IITA clones, two-step genomic prediction analyses were done. In the first step of the analyses, linear mixed models accounting for each trial's design were fitted and de-regressed BLUPs were obtained for TP1 and TP2. For TP1, we fitted the model: $y = \mathbf{X}\beta + \mathbf{Z_{clone}}c + \mathbf{Z_{range\ (loc.)}}r + \mathbf{Z_{block(range)}}b + \varepsilon$, using the *lmer* function from the *lme4* R package (R Development Core Team, 2008). In this model, $\boldsymbol{\beta}$ defined the fixed effect for the population mean and location, with **X** as the corresponding incidence matrix. The incidence matrix $\mathbf{Z_{clone}}$ and the vector *c* represented random effect for clones $c \sim \mathrm{N}(0, \mathbf{I}\sigma_c^2)$, and **I** represented the identity matrix. The range variable, which was the row or column along which plots were arrayed, was nested in location-replication and was represented by the incidence matrix $\mathbf{Z_{range(loc.)}}$ and random effects vector $r \sim \mathrm{N}(0, \mathbf{I}\sigma_r^2)$. Block effects were nested in ranges and incorporated as random term with incidence matrix $Z_{block(range)}$ and effects vector $b \sim \mathrm{N}(0, \mathbf{I}\sigma_b^2)$. Residuals ε were distributed as $\varepsilon \sim \mathrm{N}(0, \mathbf{I}\sigma_\varepsilon^2)$.

For TP2 (522 clones), we fitted the linear mixed model $\boldsymbol{y} = \mathbf{X}\beta + \mathbf{Z_{clone}}c + \mathbf{Z_{block}}b + \varepsilon$, where *y* was the vector of raw phenotypes, $\boldsymbol{\beta}$ included a fixed effect for the population mean and location with checks included as a covariate. The incidence matrix $\mathbf{Z_{clone}}$

and the vector *c* were similar for both TP1 and TP2. The blocks were also modeled with incidence matrix $\mathbf{Z_{block}}$, and $\boldsymbol{b}$ represented the random effect for the blocks. The best linear unbiased predictors (BLUPs) of the clone effect were extracted as de-regressed BLUPs following the formula proposed by Garrick et al., (2009).

$$\text{deregressed BLUP} = \frac{\text{BLUP}}{1 - \frac{\text{PEV}}{\sigma_c^2}}$$

Here, PEV represented the prediction error variances for the BLUPs and σ_c^2 was the clone variance.

For combined data set of IITA clones as a test set, we fitted the model: $y = \mathbf{X}\beta + \mathbf{Z_{clone}}c + \mathbf{Z_{rep(trial)}}b + \varepsilon$, where *y* was a vector of raw phenotypes, $\boldsymbol{\beta}$ included a fixed effect for the population mean. The incidence matrix $\mathbf{Z_{clone}}$ and the vector *c* represented random effect for clones $c \sim \mathrm{N}(0, \mathbf{I}\sigma_c^2)$ and **I** represented the identity matrix, and the replication nested in the trials was modeled with incidence matrix $\mathbf{Z_{rep(trial)}}$, with random effect $\boldsymbol{b}$ representing replications nested within trial for the first set of 27 and second set of 22 clones, evaluated with some 14 overlapping clones between the sets. The best linear unbiased predictors (BLUPs) were extracted from the model and subsequently used as the validation data for estimation of genomic prediction accuracies of CBSD for the 35 unique IITA clones. In addition, variance components were extracted from the model to compute plot based broad-sense heritability estimates.

Population structure

To assess population structure, we used the GBS markers of TP1, TP2, and the 35 IITA clones. These markers were filtered to have a minor allele frequency (MAF) ≥ 0.01 and formatted as a dosage matrix with SNP genotypes coded as -1, 0, or +1. Principal component

analysis (PCA) was done on the SNP matrix, using the *prcomp* function in R. The first two principal components (PC) were used to visualize population structure.

Cross-validation prediction accuracies for IITA clones

We estimated prediction accuracies for foliar CBSD severities evaluated at three (CBSD3s) and six (CBSD6s) months, and root severity at 12 months (CBSDRs), using a 5-fold cross validation scheme, replicated 10 times for IITA clones from a single-step genomic best linear unbiased predictor (G-BLUP) model. For each replication in the cross-validation scheme, the 35 IITA clones were randomly divided into five groups of 7 clones each (folds). Four groups at a time were used as the training population to build the prediction model, while excluding the fifth group, which was used as the model validation set. This was repeated for all the 5-folds for each of the 10 replications. Prediction accuracies were computed as the Pearson correlation coefficient between the genomic estimated breeding values predicted for the validation set and the corresponding BLUPs obtained from the first-step of the analysis for 35 unique individuals in the test set (IITA clones).

Genomic prediction of CBSD for IITA clones

We tested genomic prediction accuracies under four scenarios: (i) optimized training populations across genomic selection models (ii) optimized versus random subset training populations for G-BLUP only (iii) models with inclusion of kernels defined by chromosomes on which CBSD QTL have been found (single and multi-kernel G-BLUP models), and (iv) high and low density marker panels for G-BLUP model.

To optimize the training population, we used the selection of training population with a genetic algorithm (STPGA), *GenAlgForSubsetSelection*, from the R package STPGA (Akdemir et al., 2015). The algorithm identifies a subset of a specified size from a larger pool of potential training individuals. To do this, STPGA finds the set of individuals that minimize

the mean prediction error variance (mean PEV) expected for test set, using molecular marker data.

For STPGA training population optimization, we used the first 50 principal components (PC's) of the eigenvalue decomposition of the marker matrix as a dimension reduction approach. The pool of potential training individuals was the combined TP1 and TP2 (N = 922) described above and the target or test set were the IITA clones. We optimized 20 training populations within each size of training population specified in STPGA. In scenario (i), the optimum training populations for each training population size (100, 200, 400, 800 and full set = 922) were used to predict CBSD with four genomic prediction models namely; G-BLUP, Bayes-A, Bayes-B and Bayesian Lasso (Lorenz et al., 2011; Heslot et al., 2012).

Under scenario (ii), we tested the performance of STPGA by comparing the optimized sets from scenario (i) with random subsets of the same size. We chose to compare optimized and random sets for population of sizes of 200 and 400, based on results from analyses in scenario (i), and for each training size we compared 20 sets for both random and optimized TPs, using G-BLUP model because of it robustness and computational efficiency.

In the single kernel model, all GBS markers were fitted with one realized genomic relationship matrix K, according to the formula described by VanRaden, (2008). The relationship matrix was constructed using *A.mat* function in rrBLUP package (Endelman, 2011). The model was specified as: $y = 1_n u_0 + Zg + e,$ with $g \sim N(0, K\sigma_g^2)$ and $e \sim N\,(0, I\sigma_e^2)$, where y was the vector of de-regressed BLUPs, u_0 was an overall population mean, Z was the design matrix linking observations to genomic values, g was the vector of genomic estimated breeding values for each clone, and e was the vector of residuals. We assumed, g had a known covariance structure defined by the realized genomic relationship matrix *K*.

Previously, QTL for CBSD have been reported on chromosomes 4 and 11 (Kawuki et al., 2016; Kayondo et al., 2018). We used all the markers on the two chromosomes (Chr.4 and 11) because significant markers covered essentially the whole of Chr. 4 for CBSD6s and about half of Chr. 11 for both CBSD3s and CBSD6s (Kayondo et al., 2018). Therefore, we also fitted a multi-kernel G-BLUP model with two realized genomic relationship matrices, constructed using *A.mat* function as described above. In this model, the first genomic relationship matrix incorporated all markers from both chromosome 4 and 11, while the second genomic relationship matrix was derived from the rest of the genomic markers. The model was: y = $1_n u_0 + Zq + Zr + e$. Here, y was the vector of de-regressed BLUPs, u_0 was an overall mean, Z was the design matrix linking observations to genomic values, q was the vector of genomic values captured by combined QTL markers linked to CBSD resistance, r was the vector of genomic values captured by the remaining set of genetic markers, and e was a vector of residuals. The random genetic effects for both kernels with their variance-covariance structure K, and the residuals were assumed to be normally distributed as $q \sim N\ (0, K_q\sigma_q^2)$, $r \sim N\ (0, K_r\sigma_r^2)$ and $e \sim N\ (0, I\sigma_e^2)$.

We also fitted a multi-kernel G-BLUP model with three genomic relationship matrices, where the first and second realized genomic relationship matrices were defined by all the markers on chromosomes 4 and 11 respectively, while the third contained markers from the remaining 16 chromosomes. The model was: y = $1_n u_0 + Zp + Zs + Zr + e$. Here, y was the vector of de-regressed BLUPs, u_0 was an overall mean, Z was the design matrix linking observations to genomic values, p and s were the vectors of genomic values captured by QTL markers on chromosome 4 and 11 respectively, r was the vector of genomic values captured by the remaining set of genetic markers, and e was the vector of residuals. The random effects, including the residual-term were assumed to be normally distributed as $p \sim N\ (0, K_p\sigma_p^2)$, $\sim N\ (0, K_s\sigma_s^2)$, $r \sim N\ (0, K_r\sigma_r^2)$ and $e \sim N(0, I\sigma_e^2)$.

For both single and multi-kernel G-BLUP analyses, we used the two EMMREML functions, *emmreml* and *emmremlMultiKernel* to fit single and multi-kernel G-BLUP models respectively (Akdemir and Okeke, 2015). Lastly, we tested prediction accuracies of CBSD3s, CBSD6s and CBSDRs using high-density (WGS) imputed markers and compared that to low density (GBS) markers used in the analyses described above. For the high-density set, we fitted the single- and multi-kernel G-BLUP models described above, using training populations of 200 and 400 clones that were optimized using either the GBS or the WGS markers. Because the results of these two optimizations were quite similar, we only reported results from the GBS optimizations.

Results

Population structure, heritability and cross-validation within IITA clones

Principal component analyses on the SNP marker matrix showed no genetic differentiation among the TP1, TP2 and IITA clones. This was supported by PC1 and PC2 explaining only 8.75% and 5.69% of the total genetic variations, respectively (Figure 4. 1 and Figure S4. 1). Estimates of plot-basis broad-sense heritability (H^2) were computed for CBSD3s, CBSD6s and CBSDRs for the 35 IITA clones (Table 4. 1). Broad-sense heritability estimates spanned from 0.42 to 0.64 for CBSD3s and CBSDRs respectively. In addition to broad-sense heritability, we estimated narrow-sense heritability for IITA clones using a single step G-BLUP model.

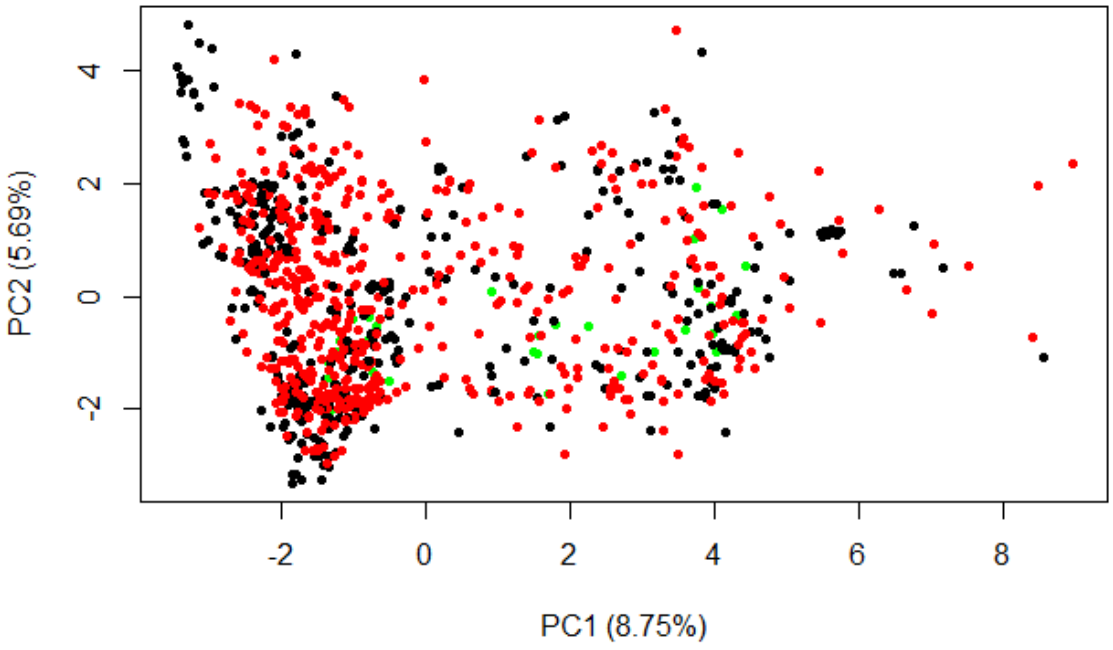


Figure 4. 2: Plot of PC1 against PC2 for Eigen value decomposition of GBS markers for IITA (**green**), NaCRRI-TP1 (**black**) and NaCRRI-TP2 (**red**) clones

Table 4. 2: Variance component and plot-basis heritability estimates for IITA clones

Source of Variation	CBSD3s	CBSD6s	CBSDRs
Clones	0.13	0.29	1.01
Reps/trial	0.01	0.00	0.00
Residuals	0.31	0.21	0.56
H^2	0.42	0.58	0.64

CBSD3s = Cassava brown streak disease severity scored at three months, CBSD6s = Cassava brown streak disease severity scored at six months, CBSDRs = Cassava brown streak disease root severity scored at 12 months, H^2 = plot-based broad-sense heritability estimates.

The lowest and highest narrow-sense heritability of 0.35 and 0.69 were recorded for CBSD3s and CBSDRs, respectively (Figure 4. 3). The average prediction accuracies from 5-fold cross-validation replicated 10 times for the IITA clones were 0.40, 0.21 and 0.08 for CBSD3s, CBSD6s and CBSDRs, respectively (Figure 4. 4). We did not do cross validation within the training set here, because the training population was previously cross-validated (Kayondo et al., 2018). Previous predictive accuracy for CBSD-related traits, had mean values across methods of 0.29 (CBSD3s), 0.40 (CBSD6s) and 0.34 (CBSDRs) for cross-validation within NaCRRI training set.

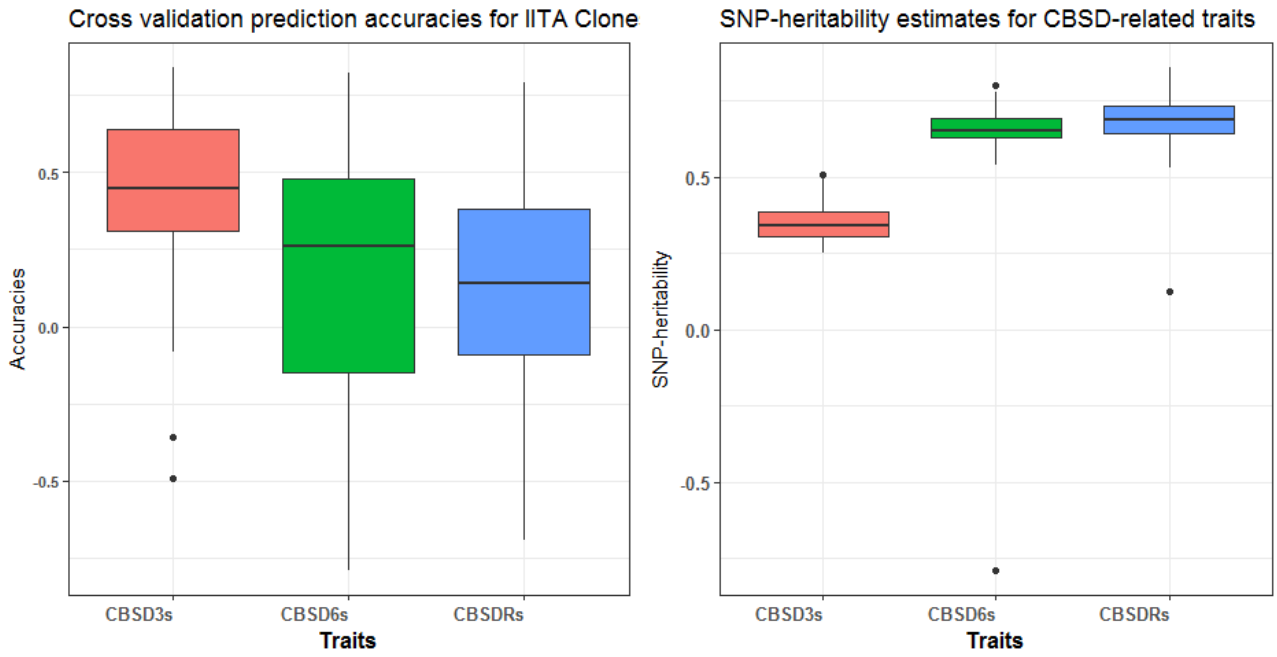


Figure 4. 5: Prediction accuracies for 5-fold and 10 reps, G-BLUP model for CBSD3s, CBSD6s and CBSDRs, and SNP heritability estimates for CBSD in 35 IITA clones

Predicting CBSD in IITA clones using Ugandan training populations

In general, the mean CBSD prediction accuracies were higher for foliar than root necrosis for the different optimized training population sizes across genomic prediction models (Table 4. 3). For CBSD3s, the prediction accuracies ranged from 0.24 (Bayes-A) to 0.36 (Bayesian Lasso). Prediction accuracies spanned from 0.14 (Bayesian Lasso) to 0.28 (G-BLUP) for CBSD6s. For CBSDRs, accuracies included negative values, ranging from -0.29 (Bayes-A) to 0.11 (Bayes-A) across different optimized training sets.

The models did not differ much in terms of their prediction accuracies for three traits (CBSD3s, CBSD6s and CBSDRs) across optimized training populations of 100, 200, 400, 800, and full set of 922 clones. Surprisingly, Bayesian Lasso consistently had higher prediction accuracies than the other three prediction models (G-BLUP, Bayes-A and Bayes-B) for CBSD3s across the optimized TP sizes, but performed worse than those three models for CBSD6s across optimized TPs (Table 4. 4).

Prediction accuracies for optimized training populations across the four models tested increased from 100 to 400 for CBSD3s to attain a plateau and declined as the optimized training population was increased to 800 and the full set of 922 clones. However, no clear trend in prediction accuracies were observed for CBSDRs for the different sizes of optimized training population (Table 4. 5).

Table 4. 6: Average prediction accuracies (r) for four optimized subsets of TPs and full set across genomic prediction models

TP Size	G-BLUP			BAYES-A			BAYES-B			BAYESIAN LASSO		
	CBSD3s	CBSD6s	CBSDRs	CBSD3s	CBSD6s	CBSDRs	CBSD3s	CBSD6s	CBSDRs	CBSD3s	CBSD6s	CBSDRs
TP100	0.27^{ns}	0.23^{ns}	-0.10^{ns}	0.26^{ns}	0.22^{ns}	-0.19^{ns}	0.30*	0.23^{ns}	-0.03^{ns}	0.33*	0.19^{ns}	-0.07^{ns}
TP200	0.27^{ns}	0.28^{ns}	-0.03^{ns}	0.26^{ns}	0.26^{ns}	-0.29^{ns}	0.27^{ns}	0.26^{ns}	0.07^{ns}	0.34*	0.22^{ns}	0.06^{ns}
TP400	0.32*	0.19^{ns}	-0.01^{ns}	0.32*	0.18^{ns}	-0.19^{ns}	0.32*	0.17^{ns}	-0.09^{ns}	0.36*	0.14^{ns}	-0.08^{ns}
TP800	0.31*	0.26^{ns}	0.06^{ns}	0.29^{ns}	0.25^{ns}	-0.13^{ns}	0.29^{ns}	0.23^{ns}	-0.04^{ns}	0.31*	0.17^{ns}	-0.01^{ns}
TP922	0.30*	0.25^{ns}	0.05^{ns}	0.24^{ns}	0.21^{ns}	0.11^{ns}	0.30*	0.26^{ns}	-0.09^{ns}	0.31*	0.15^{ns}	-0.04^{ns}

CBSD3s = Cassava brown streak disease severity scored at three months, CBSD6s = Cassava brown streak disease severity scored at six months, CBSDRs = Cassava brown streak disease root severity scored at 12 months; TP100, TP200, TP400, TP800 and TP922 = Optimized training populations of size 100, 200, 400, 800 and a full set of 922 clones, ns = non-significant prediction accuracies (r), * accuracy significantly different from zero ($P \leq 0.05$).

We compared CBSD prediction accuracies from random and optimized training populations of size 200 and 400 clones using the G-BLUP model. We chose these two sample sizes because they maximized prediction accuracies for CBSD3s and CBSD6s (Table 4. 7). For both 200 and 400 clones, the prediction accuracies were higher for optimized training sets than for the random subsets for CBSD3s and CBSD6s (Figure 4. 6 and Table S4. 6). For example, at training population size of 200, the mean prediction accuracies for CBSD3s and

CBSD6s were 0.27 and 0.28 compared to 0.11 and -0.01 for the corresponding random subsets. Similarly, at training population size of 400 clones, the mean prediction accuracies for CBSD3s and CBSD6s were 0.32 and 0.19 relative to 0.10 and 0.04 for the random subsets (Figure 4. 7 and Table S4. 7). We observed markedly lower standard errors as measures of variation in prediction accuracies across the traits for the optimized training populations, compared to the random subsets (Figure 4. 8). However, no strong differences were observed for CBSDRs (Figure 4. 9).

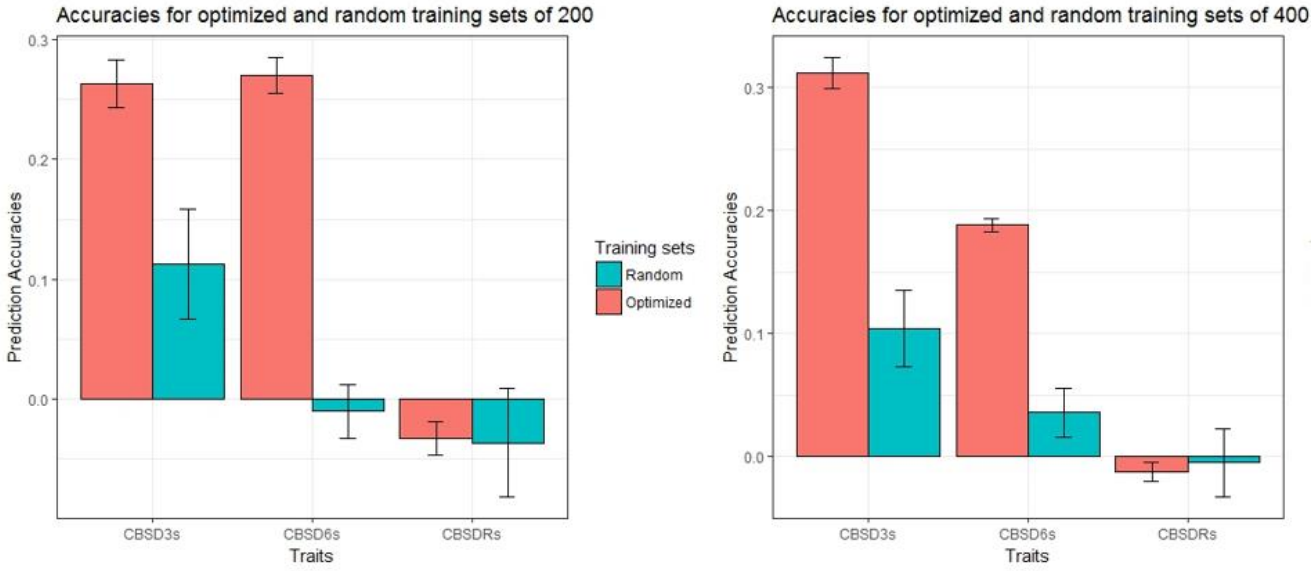
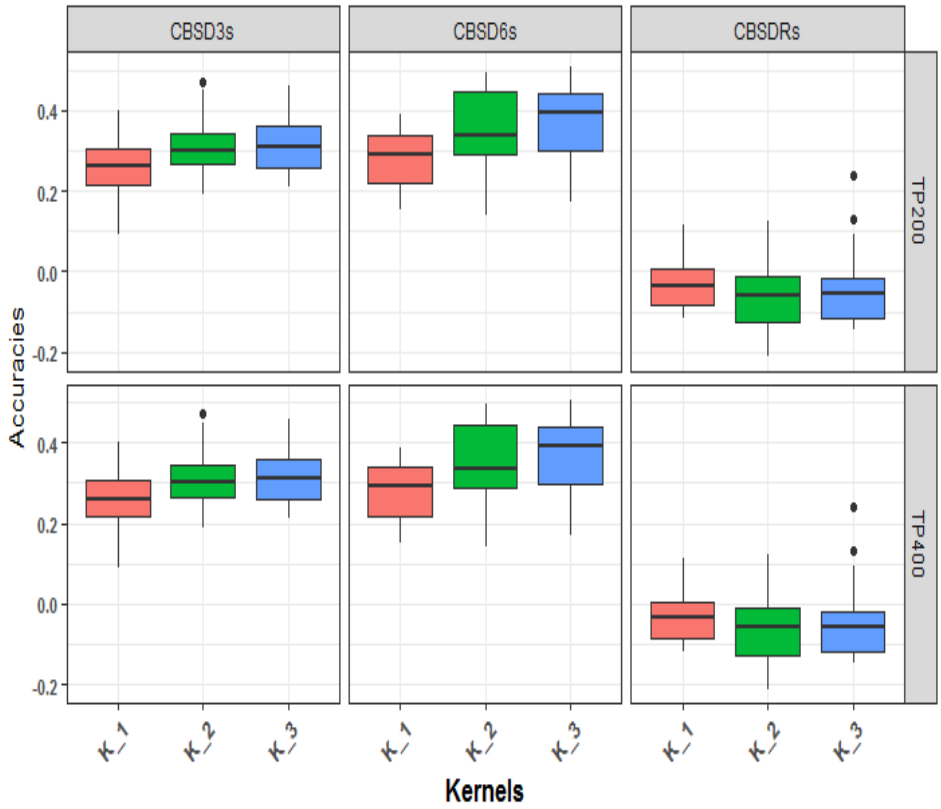


Figure 4. 10: Prediction accuracies and the standard error bars for 20 replications of optimized and random training population size of 200 and 400

Accounting for CBSD QTL with chromosome-specific effects or kernels

In general, foliar CBSD prediction accuracies for training population size of 200 and 400 were higher for multi-kernel models (K_2 and K_3) with separate kernels fitted for CBSD QTL chromosome markers than single kernel (K_1) G-BLUP models (Figure 4. 11). Prediction accuracies for CBSD3s increased from 0.27 for the single kernel G-BLUP, termed as "K_1" model to 0.31 for two-kernel G-BLUP model referred to as "K_2", and to 0.32 for the three-kernel model referred to as "K_3" in the optimized TPs of 200 clones (Figure 4. 12 and Table S4. 8). Similarly, for CBSD6s, prediction accuracies increased from 0.28 for single kernel G-

BLUP model to 0.37 with three-kernels (Figure 4. 13 and Table S4. 8). No such increase was observed for CBSDRs. Notably, the mean prediction accuracies for CBSD3s and CBSD6s from multi-kernel G-BLUP models were statistically significantly different ($P \leq 0.05$) from zero. Nevertheless, no differences were observed for CBSDRs prediction accuracies between single- and multi-kernel G-BLUP models at TP size of 200.



K_1= Single kernel G-BLUP model

K_2= Multi-kernel G-BLUP, the first kernel is defined by combined markers from chromosomes 4 and 11. The second kernel is defined by the remaining markers

K_3= Multi-kernel G-BLUP, the first and second kernels are defined by markers from chromosomes 4 and 11 respectively, and third kernel is defined by the remaining markers.

Figure 4. 14: G-BLUP model to compare prediction accuracies for varying number of kernels for CBSD measured at 3, 6 and 12 MAP for size of TP 400 and 200

For the optimized training population size of 400, a similar trend of increased prediction accuracies was observed from single- to multi-kernel G-BLUP models for both CBSD3s and CBSD6s. The mean prediction accuracies for CBSD6s were not significantly different from zero. Prediction accuracies did not vary much for CBSDRs between single- and multi-kernel G-BLUP models (Figure 4. 15 and Table S4. 9).

Comparing prediction accuracies for high (WGS) and low (GBS) density markers

Single kernel G-BLUP prediction accuracies for CBSD3s and CBSDRs were higher for WGS, than GBS markers for both optimized training population sizes of 200 and 400 clones (Figure 4. 16). For CBSD6s, however, predictions accuracies were lower for high density (WGS) at both training population sizes. For single kernel G-BLUP, prediction accuracies for CBSD3s and CBSDRs increased from 0.27 to 0.35, and -0.03 to 0.18 from low to high density marker sets, respectively at the optimized training population size of 200 clones (Table S4. 11). Similarly, predictions accuracies for CBSD3s and CBSDRs increased from 0.32 to 0.39, and -0.01 to 0.16 for low density (GBS) and high density (WGS) imputed markers respectively for the training populations of 400 clones (Figure 4. 17 and Table S4. 12).

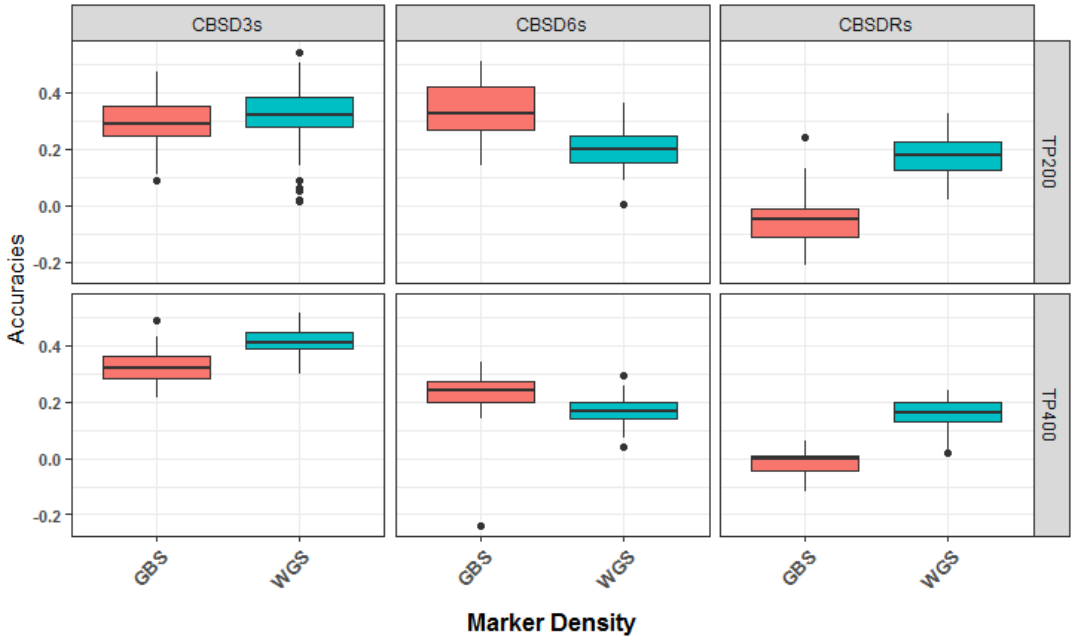


Figure 4. 18: Comparison of prediction accuracies for the CBSD-related traits under high density, whole genome sequence imputed (WGS) and low density genotyping-by-sequencing (GBS) markers for optimized training population sizes of 200 and 400 clones using single kernel

Fitting multi-kernel G-BLUP models including kernels defined by markers on CBSD QTL chromosomes 4 and 11 for high density markers did not always increase prediction accuracies (Tables S4. 11 and S4. 12). The highest prediction accuracy of 0.44 (CBSD3s) was

recorded from the multi-kernel G-BLUP model, fitted with high density (WGS) imputed marker for the optimized training population of 400 clones (Table S4. 12). In other cases, prediction accuracies actually dropped from single- to multi-kernel models. For example, prediction accuracy for CBSD3s dropped from 0.35 for single kernel to 0.32 for multi-kernel (three kernels) at the training population size of 200 clones (Table S4. 11). Overall, the prediction accuracies for high and low-density marker sets were similar between the multi-kernel models regardless of the optimized training population size (Tables S4. 11 and S4. 12).

Discussion

Cassava brown streak disease (CBSD) caused by Uganda cassava brown streak virus (UCBSV) and cassava brown streak virus (CBSV) has continued to be a major threat to cassava productivity in southern, eastern and central parts of Africa. Recently, CBSD causing viruses were declared the leading biological enemy to cassava productions in CBSD endemic zones of Sub-Sahara Africa (Legg et al., 2014). Concerted efforts such as quarantine, disease surveillance, and breeding for resistance have taken center-stage to prevent further spread of CBSD to W. Africa, especially Nigeria, the world's largest producer and consumer of cassava. In this chapter, we leveraged genome-wide prediction approaches as a potential means to enable pre-emptive breeding for CBSD resistance in W. Africa.

Impact of different sizes of optimized training population across models

For optimized training populations of 100, 200, 400, 800 and 922 clones, the highest prediction accuracies were observed at the training population sizes of 200 (G-BLUP) and 400 (Bayes-B) clones for CBSD6s and CBSD3s respectively. Our findings were similar to that of Wolfe et al. (2017), where prediction accuracy of 0.37 for CMD was observed for both the smallest and largest optimized training sizes of 300 and 900, respectively in cross-population prediction, suggesting that accuracies similar to that of the full set can be obtained with a small but carefully selected TP in relation to the test set.

Overall, the cross-population prediction accuracies for IITA clones, based on optimized training populations and various prediction models (spanning 0.24 to 0.36), were comparable for CBSD3s to those reported previously for the cross-validation within NaCRRI training set, ranging from 0.27 to 0.32 (Kayondo et al., 2018). In contrast, for CBSD6s, our prediction accuracies (0.14 to 0.28) were lower than accuracies reported by Kayondo et al. (2018), which ranged from 0.40 to 0.42. The similarity in foliar CBSD prediction accuracy for CBSD3s,

indicates some genetic signal for CBSD foliar symptom expression for IITA clones was captured by optimal NaCRRI training subsets. Unfortunately, our cross-population prediction accuracies for CBSDRs for optimized TPs were generally lower than the accuracies reported for cross-validation within NaCRRI training population for CBSDRs. In part, the negative prediction accuracies for CBSDR could be explained by G x E interaction for TP1 and TP2 (Table S4. 13 and Figure S4. 4). However, the G x E variances relative to genetic variances were low and therefore unlikely to explain fully the poor prediction accuracies observed for CBSD root necrosis in W. African clones**.** The low prediction accuracy for root necrosis suggests the need to phenotype clones of W. African descent (i.e., belonging to the W. African subpopulation) in E. Africa, and subsequently to use that data for predicting CBSD resistance in W. African clones. One option would be to send many W. African clones to E. Africa as tissue culture plantlets. As observed in this study, cost and mortality are high for this option. Another possibility would be to send botanical seeds of W. African clones to E. Africa for evaluation. As shown in chapter two that root necrosis scores on seedlings have high genetic correlation with root necrosis scores on clonally propagated plants. Thus, evaluation of seedlings of W. African origin could provide a progeny test of W. African clones and the resulting breeding values could be used to train prediction models for W. Africa.

We did not observe consistent superior performance for any of the prediction models that we tested or for any of the CBSD traits analyzed. Several studies have reported similar results in that most prediction models perform similarly (Jannink et al., 2010; Heslot et al., 2012; Roorkiwal et al., 2016). Even though the models tested in the present study assumed different distributions of marker effects (Meuwissen et al., 2001; Lorenz et al., 2011), their similarity in prediction accuracies could be interpreted as approximation to optimal genomic prediction models, where all the models capture the same or similar QTL effects across the

genome (Su et al., 2014). In such a situation, the choice of GS model would be less important than the actual design of the training population for across-population predictions.

Comparison of prediction accuracies for random and optimized training populations

Prediction accuracies can be improved by targeting more informative individuals in the reference panel used to generate the predictions and this has been demonstrated in several crop species (Rincent et al., 2012; Akdemir et al., 2015). In general, we observed higher prediction accuracies for CBSD3s and CBSD6s from optimized compared with randomly selected training set of the same size. For example, at TP size of 200 clones, our prediction accuracies for CBSD3s was 0.27 with the optimized compared to 0.11 from the random subset. Similar findings were made by Wolfe et al. (2017), where STPGA-optimized training populations performed better than random subsets for a number of important cassava traits, including dry matter content (DMC), harvest index (HI), mean cassava mosaic disease and plant vigor. Our results, therefore, serve to further stress the importance of training population optimization for cross-population prediction.

Weighting prior biological information for CBSD prediction across population

Studies have shown increased prediction accuracies with inclusion of prior QTL information in genomic prediction models. For example, a study by Haffliger, (2016) for reproductive traits in Swiss pig breeds revealed a significant increase in prediction accuracy for piglets when previously detected reproductive trait QTL markers were included in the prediction model. From the training population used in this study, two recent studies identified CBSD QTL on chromosomes 4 for CBSD3s and CBSD6s, and 11 for CBSD6s and CBSDRs (Kawuki et al., 2016; Kayondo et al., 2018). In addition, bi-parental mapping studies have had similar results (Masumba et al., 2017; Nzuki et al., 2017). In an attempt to improve across-population prediction accuracies for CBSD symptoms, we chose to directly model the

Chromosome 4 and 11 (Chr.4 and Chr.11) QTLs by incorporating random effects for the markers on those chromosomes into our prediction. Prediction accuracies increased for CBSD3s and CBSD6s, but not for CBSDRs. The benefit was greatest for prediction accuracy of CBSD6s which increased by 9% , when three realized relationship matrices (Chr. 4 + Chr. 11 + the rest, optimized set of 200) were modeled. Although the percentage increase in prediction accuracies was less for the optimized TPs of 400 clones, we still observed increased prediction accuracies for CBSD6s, again when three relationship matrices were fitted. We observed a much higher increase in prediction accuracies for G-BLUP models that including the CBSD QTL as separate random effects, compared to the only marginal increase in prediction accuracies of 1.7% for CBSD3s and 2.5% for CBSDRs reported previously (Lozano et al., 2017).

The higher prediction accuracies we observed by accounting for the CBSD QTL suggests that the development of genomic resources for cassava (Prochnik et al., 2012), the identification of QTL by GWAS (Wolfe et al., 2016; Lozano et al., 2017; Kayondo et al., 2018) and candidate genes by bioinformatics (Lozano et al., 2015) can provide benefits for genomic prediction, particularly in across population prediction scenarios.

Comparing prediction accuracies of high and low density marker panels

In the present study, prediction accuracies for CBSD3s and CBSDRs were 8% and 18% higher for high density (WGS) imputed markers than low density (GBS) markers from single kernel G-BLUP model for the optimized training population of size of 200 clones (Table S4. 11). Several studies have demonstrated increased prediction accuracies as a function of increase marker density (Peixoto et al., 2016; Wang et al., 2017). In a recent study, using NaCRRI training population, prediction for CBSD-related traits, in a single kernel G-BLUP model was not improved by whole-genome imputation (Lozano et al., 2017). On the other hand, in a simulation study for across population genomic prediction in dairy cattle, De Roos

et al., (2009) reported higher prediction accuracies, similar to the improvement we observed for CBSD3s and CBSDRs, when more markers were included in the model. The study concluded that the reliability of genomic predictions across populations is determined by the consistency of marker–QTL allelic phase between the populations. The more diverged the populations are, the denser the markers must be to ensure preservation of marker–QTL phase across the populations. Increased prediction accuracies for CBSD3s and CBSDRs in this study, could therefore be a result of whole genome sequence imputed markers more reliably capturing the correct marker-CBSD QTL phase across the two populations. Since the only additional cost incurred in generating WGS imputed markers is the bioinformatics to generate the imputed markers in our case, we believe that imputing the GBS markers to reasonable levels using bioinformatics would benefit even poorly resourced breeding programs.

Conclusion

We have presented the first empirical validation of genomic prediction for cassava brown streak disease across populations. Based on our results, training population optimization provided a benefit of increased prediction accuracies over random subset and full set of training population for foliar cassava brown streak disease. More importantly, inclusion of prior CBSD QTL information in our genomic prediction models reasonably increased foliar CBSD prediction for W. African clones. Furthermore, whole genome sequence imputed markers increased prediction accuracies for CBSD3s and CBSDRs. Future efforts to better predict CBSD resistance in W. Africa clones could focus initially on testing progeny from W. African germplasm, and later use the progeny evaluation data to train CBSD prediction models in W. African. Lastly, further research should target a much larger number of W. African test clones than we used in the current study.

Acknowledgement

We acknowledge the Bill and Melinda Gates Foundation, and the Department for International Development of the United Kingdom for funding this work through the "Next Generation Cassava Breeding Project". Thanks to the technical field staff of NaCRRI Cassava-breeding program for generating the phenotypic data and the team from biosciences laboratory for the DNA extraction work. We also thank the International Institute of Tropical Agriculture (IITA), Ibadan, Nigeria for sharing their germplasm with us for this study. Great thanks to Cornell University Institute of Genomic diversity (IGD) for genotyping the clones

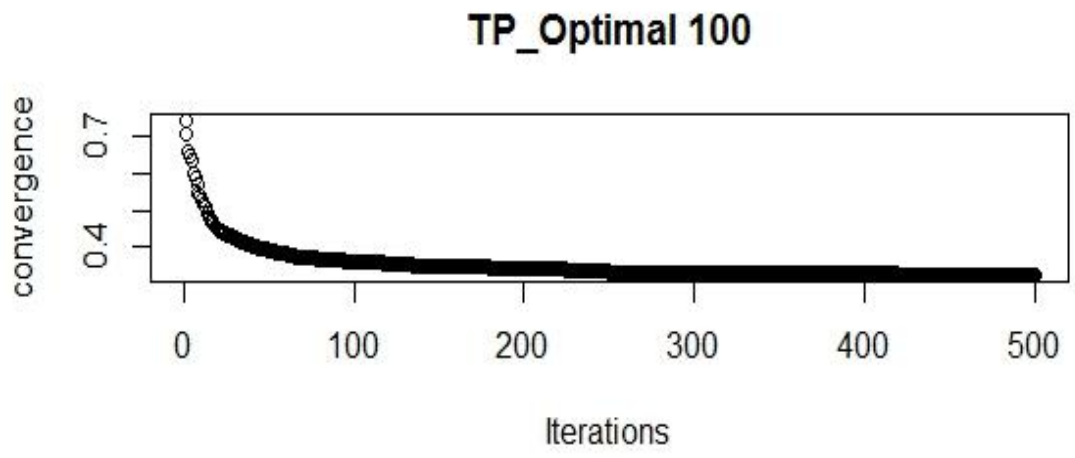


Figure S4. 2: STPGA model convergence for optimized training population of 100 clones

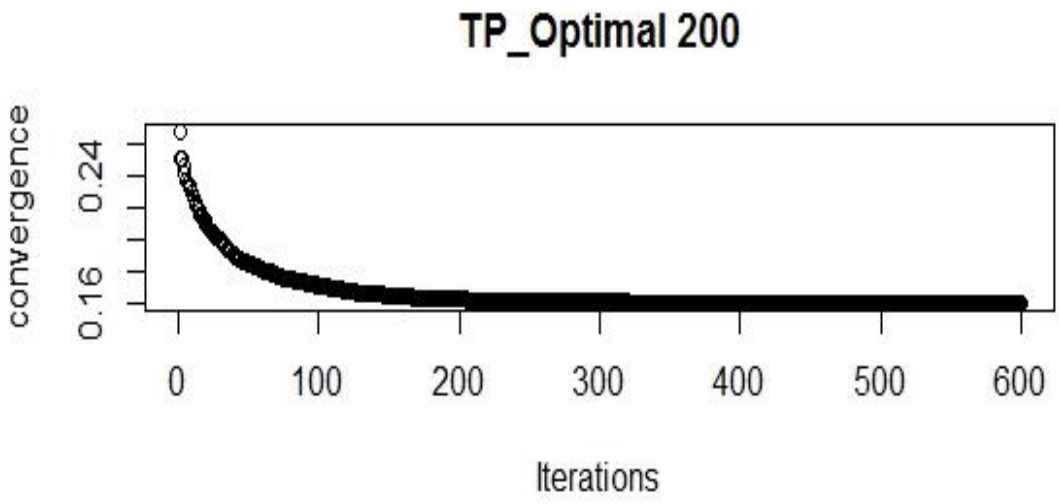


Figure S4. 3: STPGA model convergence for optimized training population of 200 clones

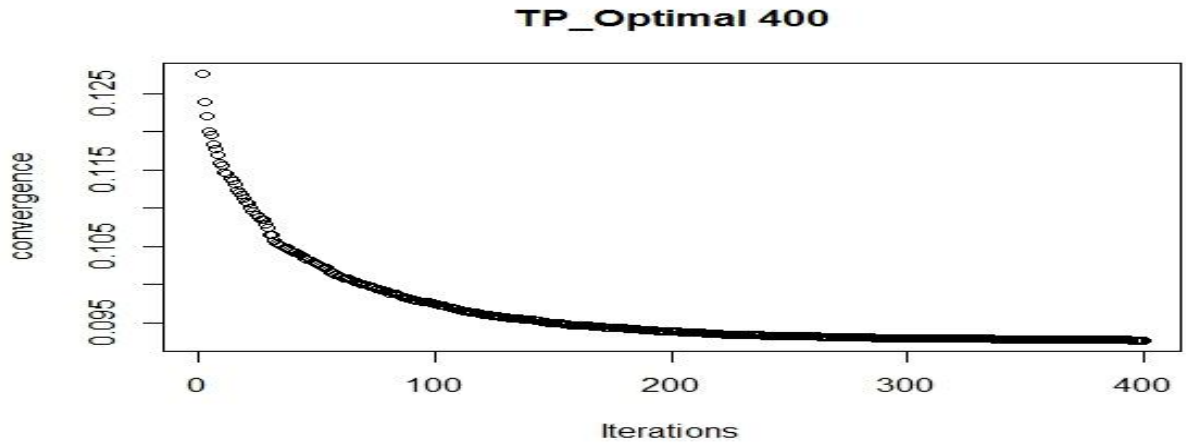


Figure S4. 4: STPGA model convergence for optimized training population of 400 clones

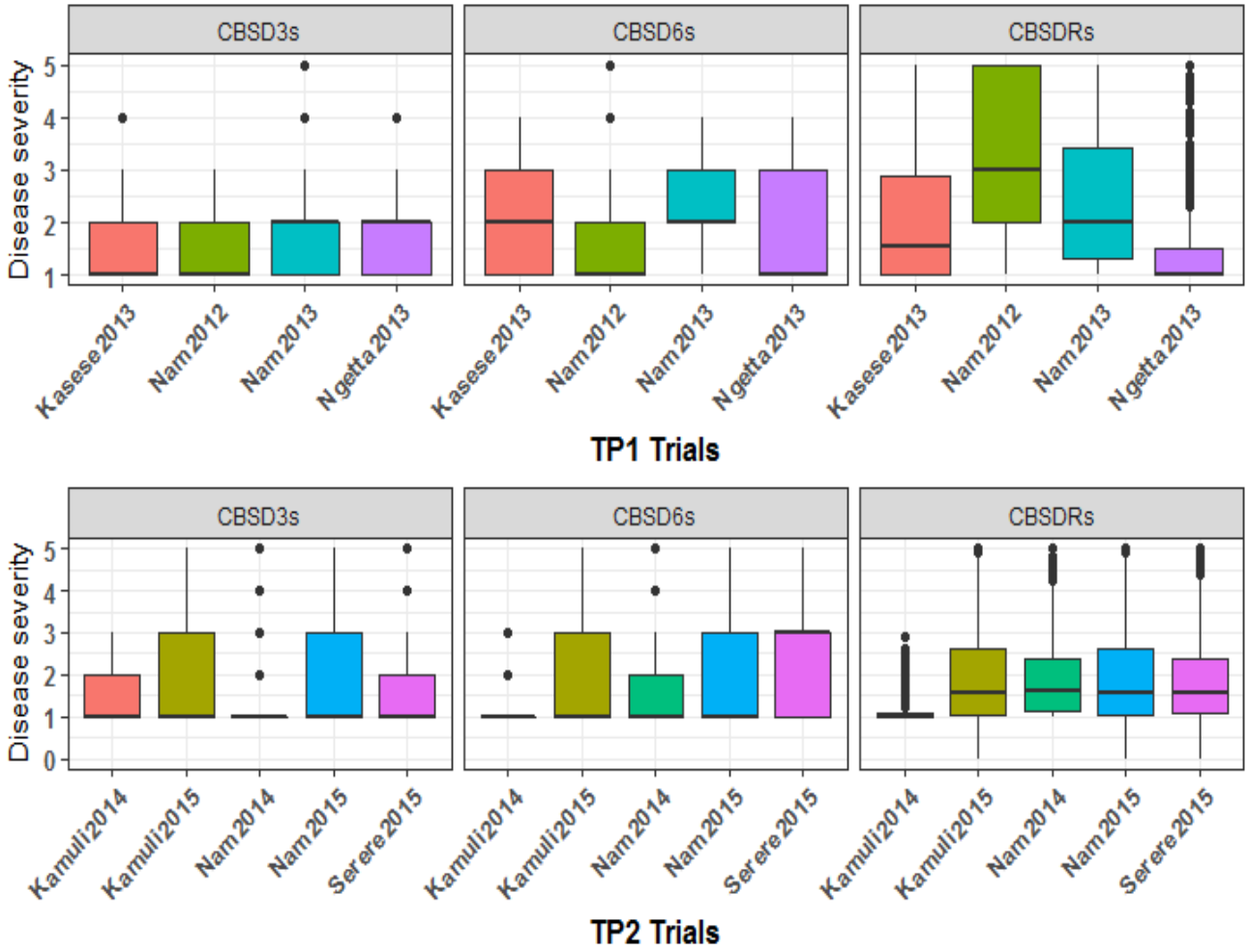


Figure S4. 5: Boxplot showing the phenotypic distribution of two training sets (TP1 and TP2) for the three disease traits.

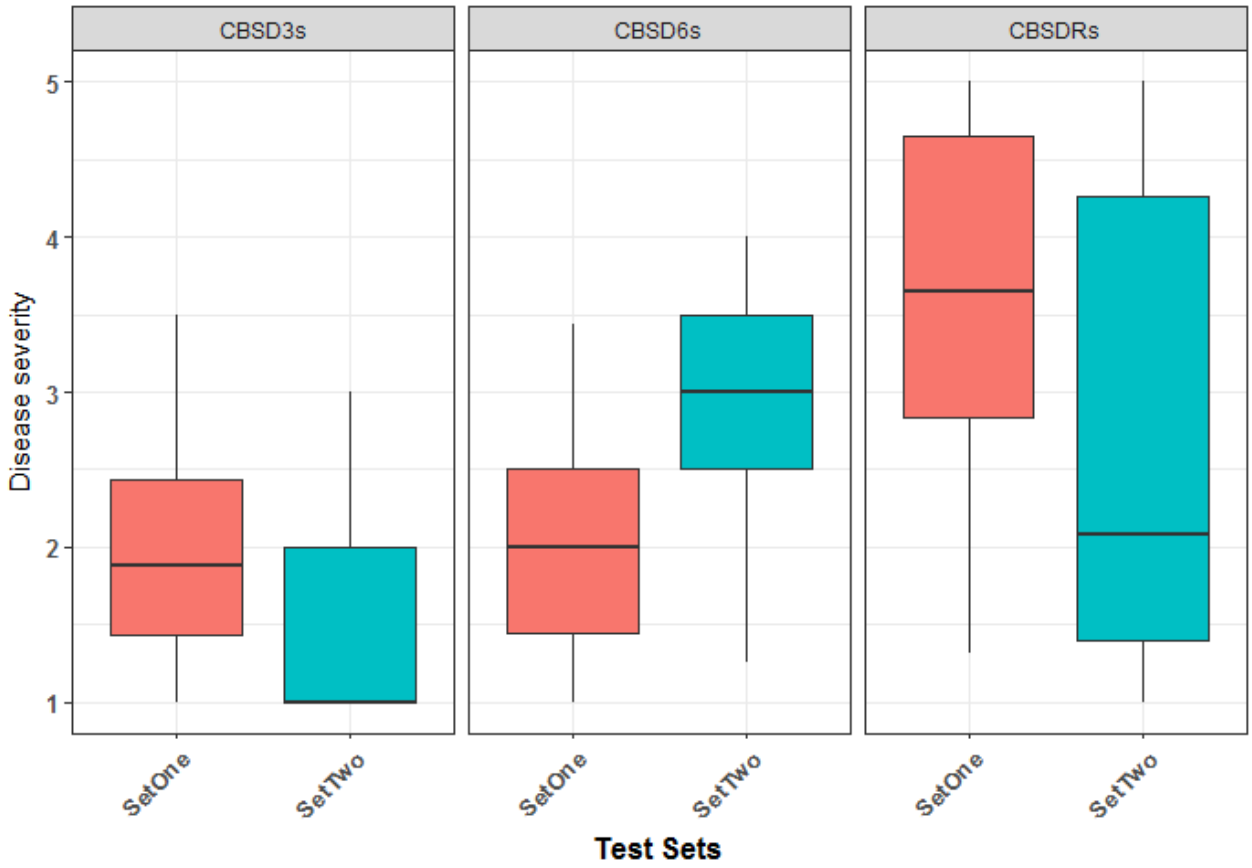


Figure S4. 6: Boxplot showing the phenotypic distribution of two sets of W. African clones for the three disease traits.

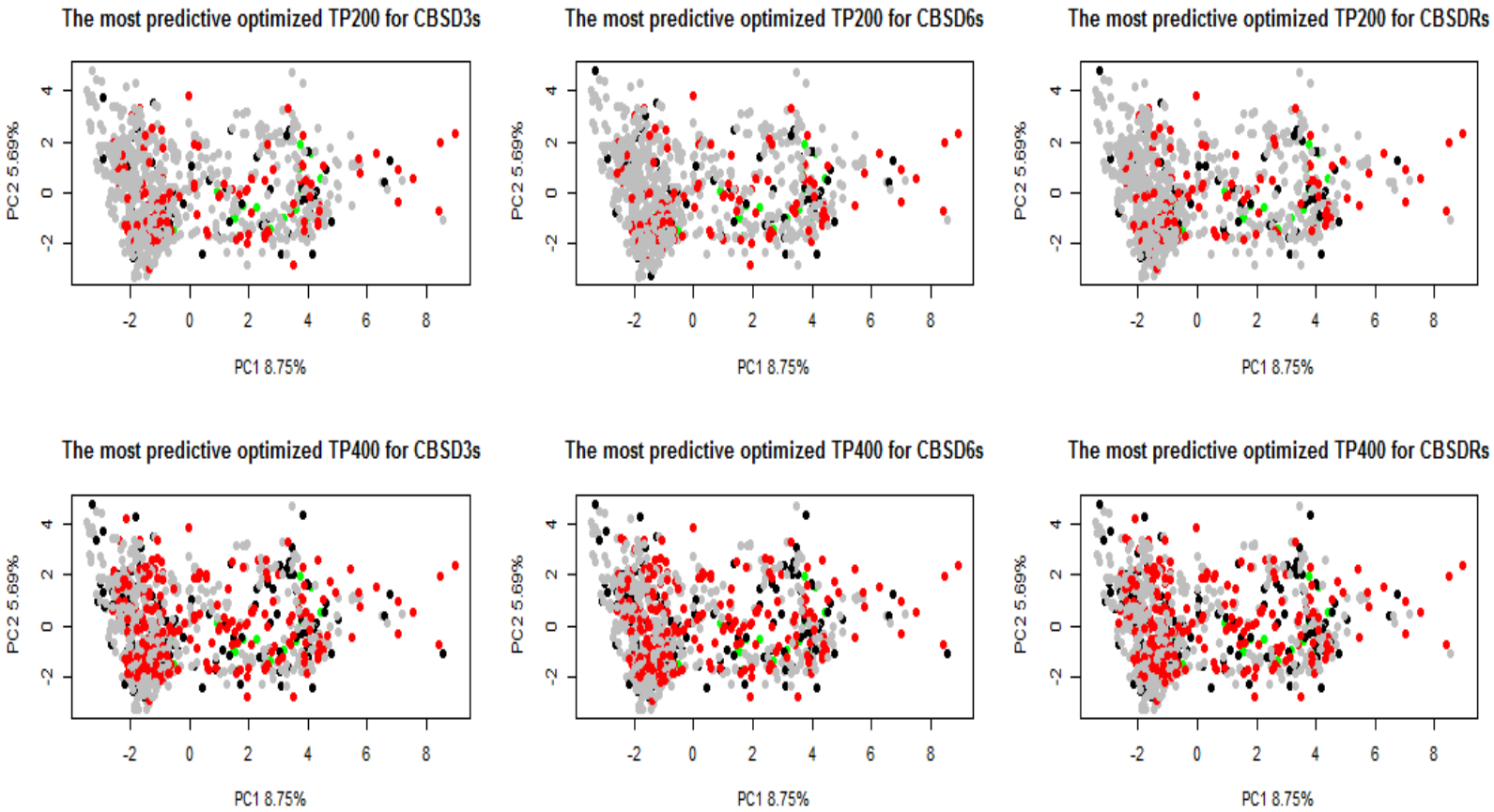


Figure S4. 7: Plot of PC1 against PC2 for the most predictive optimized training size of 200 and 400 for TP1 (**Black**) and TP2 (**Red**) as well as the unselected TP1+TP2 (**Grey**) and the IITA test set (**Green**) for the CBSD3s, CBSD6s and CBSDRs

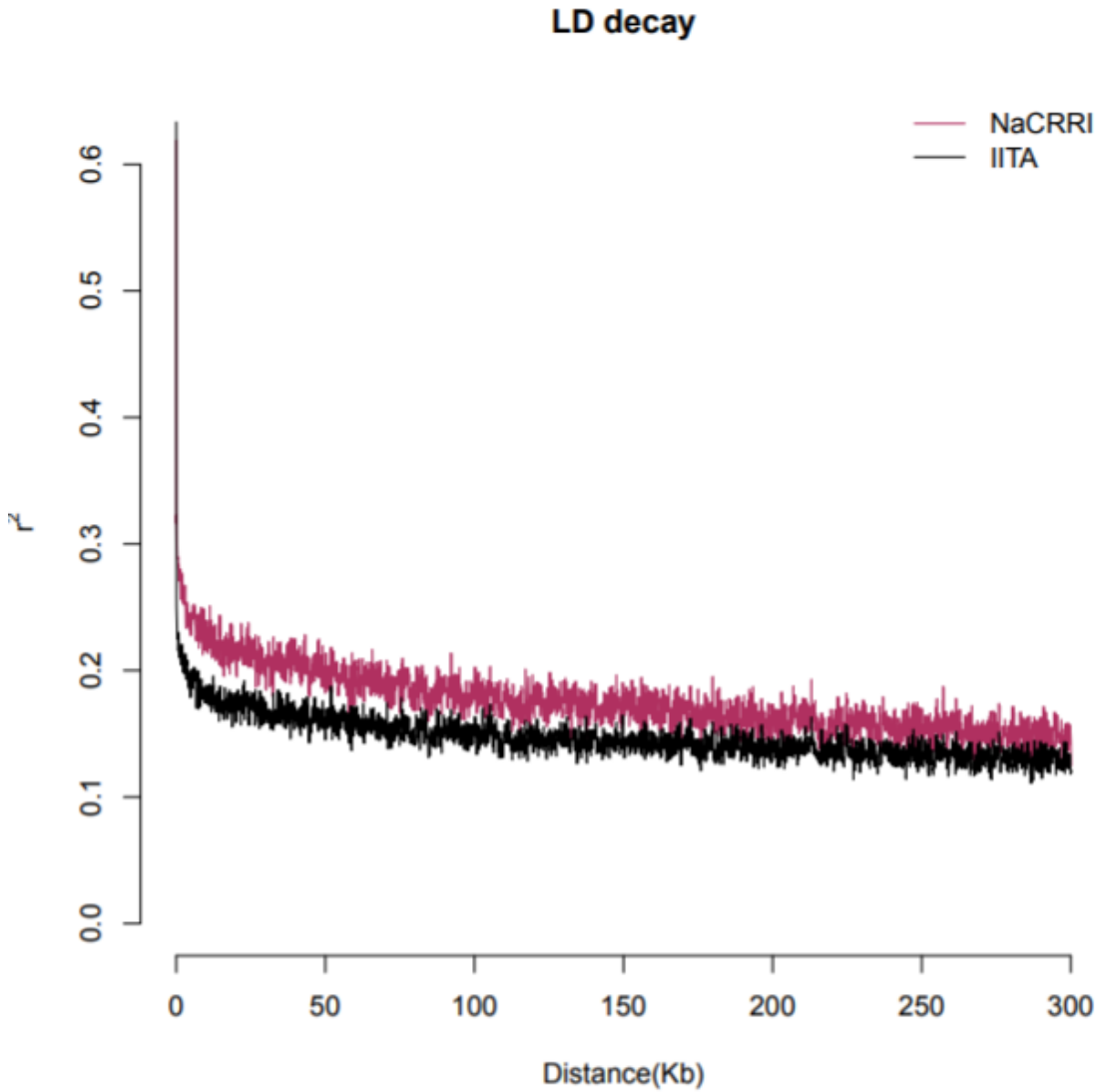


Figure S4. 8: Linkage disequilibrium (LD) decay measured as the r^2 values of pair-wise relationship among the markers along the chromosomes

Table S4. 1: Prediction accuracies for optimized training population size of 100 for combined TPs (TP1 and TP2)

Training Size of 100	G-BLUP			Bayes-A			Bayes - B			Bayesian Lasso		
	CBSD3s Pre.acc	CBSD6s Pre.acc	CBSDRs Pre.acc	CBSD3s Pre.acc	CBSD6s Pre.acc	CBSDRs Pre.acc	CBSD3s Pre.acc	CBSD6s Pre.acc	CBSDRs Pre.acc	CBSD3s Pre.acc	CBSD6s Pre.acc	CBSDRs Pre.acc
TP_1	0.21	0.21	-0.09	0.22	0.23	0.03	0.27	0.24	-0.13	0.31	0.21	-0.15
TP_2	0.36	0.22	-0.04	0.42	0.20	-0.30	0.43	0.18	-0.04	0.36	0.15	0.15
TP_3	0.09	0.12	0.00	0.09	0.11	-0.28	0.07	0.15	0.02	0.18	0.15	-0.06
TP_4	0.15	0.16	-0.16	0.14	0.13	-0.43	0.20	0.17	0.03	0.14	0.14	-0.01
TP_5	0.21	0.22	-0.07	0.23	0.19	-0.11	0.24	0.21	-0.05	0.29	0.16	0.04
TP_6	0.22	0.23	-0.05	0.19	0.20	-0.22	0.22	0.16	0.02	0.27	0.16	-0.07
TP_7	0.44	0.29	-0.19	0.38	0.25	-0.18	0.43	0.35	0.09	0.41	0.27	-0.07
TP_8	0.43	0.16	-0.29	0.38	0.15	-0.06	0.41	0.16	-0.20	0.45	0.13	-0.22
TP_9	0.27	0.28	-0.17	0.31	0.33	-0.31	0.31	0.32	0.06	0.39	0.23	-0.06
TP_10	0.11	0.19	-0.16	0.13	0.20	-0.24	0.18	0.20	-0.08	0.17	0.14	-0.19
TP_11	0.37	0.31	0.10	0.38	0.24	-0.24	0.49	0.32	0.08	0.39	0.19	-0.05
TP_12	0.27	0.19	0.03	0.32	0.15	-0.17	0.32	0.18	-0.01	0.31	0.17	-0.02
TP_13	0.15	0.28	-0.07	0.23	0.29	-0.32	0.23	0.27	-0.01	0.28	0.18	-0.08
TP_14	0.35	0.36	-0.11	0.34	0.38	-0.22	0.33	0.38	0.09	0.35	0.35	-0.03
TP_15	0.08	0.27	-0.07	0.05	0.28	-0.24	0.15	0.27	-0.11	0.24	0.22	-0.19
TP_16	0.23	0.27	-0.18	0.24	0.26	-0.07	0.27	0.25	0.06	0.40	0.19	-0.03
TP_17	0.33	0.20	-0.01	0.33	0.18	-0.21	0.31	0.18	-0.18	0.40	0.17	-0.14
TP_18	0.45	0.30	-0.08	0.37	0.29	-0.18	0.43	0.22	-0.13	0.50	0.23	-0.18
TP_19	0.23	0.24	-0.18	0.10	0.27	-0.20	0.26	0.26	-0.05	0.28	0.19	-0.05
TP_20	0.37	0.14	-0.29	0.31	0.14	0.16	0.39	0.09	-0.08	0.43	0.11	0.00
Mean Pre.acc	**0.27**	**0.23**	**-0.10**	**0.26**	**0.22**	**-0.19**	**0.30**	**0.23**	**-0.03**	**0.33**	**0.19**	**-0.07**

Table S4. 2: Prediction accuracies for optimized training population size of 200 for combined TPs (TP1 and TP2)

Train. Size 200	G-BLUP			Bayes-A			Bayes-B			Bayesian Lasso		
	M3CBSDs	M6CBSDs	CBSDRs	M3CBSDs	M6CBSDs	CBSDRs	M3CBSDs	M6CBSDs	CBSDRs	M3CBSDs	M6CBSDs	CBSDRs
TP_1	0.18	0.33	-0.02	0.20	0.31	-0.29	0.37	0.32	0.18	0.31	0.28	0.14
TP_2	0.27	0.38	0.06	0.33	0.36	-0.28	0.30	0.43	0.13	0.36	0.34	0.13
TP_3	0.09	0.34	-0.03	0.10	0.34	-0.20	0.09	0.39	0.10	0.21	0.20	0.12
TP_4	0.39	0.29	-0.05	0.40	0.32	-0.29	0.38	0.29	0.14	0.42	0.27	0.12
TP_5	0.40	0.25	-0.04	0.24	0.27	-0.37	0.42	0.18	-0.10	0.41	0.21	-0.09
TP_6	0.36	0.27	-0.08	0.36	0.24	-0.08	0.43	0.30	-0.03	0.38	0.18	-0.05
TP_7	0.28	0.29	-0.03	0.24	0.25	-0.43	0.30	0.25	0.15	0.36	0.21	0.18
TP_8	0.29	0.20	0.01	0.29	0.16	-0.47	0.42	0.22	-0.03	0.39	0.22	0.01
TP_9	0.26	0.36	-0.12	0.29	0.36	-0.22	0.20	0.31	0.13	0.35	0.29	0.11
TP_10	0.38	0.37	0.07	0.42	0.35	-0.16	0.33	0.34	0.00	0.43	0.22	-0.03
TP_11	0.26	0.29	0.11	0.21	0.27	-0.35	0.27	0.28	0.12	0.30	0.23	0.09
TP_12	0.20	0.31	-0.05	0.27	0.31	-0.46	0.22	0.30	0.22	0.24	0.22	0.23
TP_13	0.15	0.14	0.00	0.23	0.15	-0.35	0.12	0.13	-0.12	0.29	0.13	-0.07
TP_14	0.22	0.22	-0.02	0.23	0.21	-0.22	0.28	0.24	0.10	0.23	0.21	0.12
TP_15	0.28	0.31	-0.09	0.23	0.23	-0.39	0.29	0.17	0.10	0.31	0.23	0.12
TP_16	0.25	0.17	-0.09	0.27	0.09	-0.11	0.11	0.12	-0.09	0.29	0.13	-0.07
TP_17	0.11	0.19	-0.12	0.07	0.18	-0.25	0.08	0.14	-0.02	0.27	0.15	-0.07
TP_18	0.25	0.23	-0.09	0.27	0.19	-0.30	0.19	0.27	0.13	0.32	0.19	0.08
TP_19	0.27	0.21	-0.09	0.30	0.19	-0.31	0.33	0.20	0.10	0.36	0.21	0.05
TP_20	0.38	0.27	0.02	0.25	0.33	-0.26	0.37	0.35	0.13	0.31	0.26	0.15
Mean Pre.acc	**0.27**	**0.28**	**-0.03**	**0.26**	**0.26**	**-0.29**	**0.27**	**0.26**	**0.07**	**0.34**	**0.22**	**0.06**

Table S4. 3: Prediction accuracies for optimized training population size of 400 for combined TPs (TP1 and TP2)

Train.pop	G-BLUP			Bayes-A			Bayes-B			Bayesian Lasso		
	M3CBSDs	M6CBSDs	CBSDRs	M3CBSDs	M6CBSDs	CBSDRs	M3CBSDs	M6CBSDs	CBSDRs	M3CBSDs	M6CBSDs	CBSDRs
TP_1	0.40	0.18	0.01	0.40	0.18	-0.30	0.42	0.14	-0.12	0.46	0.13	-0.11
TP_2	0.28	0.19	0.06	0.26	0.21	-0.11	0.32	0.16	-0.05	0.34	0.12	-0.07
TP_3	0.32	0.16	0.01	0.32	0.16	-0.10	0.33	0.12	-0.11	0.38	0.12	-0.10
TP_4	0.34	0.21	-0.05	0.33	0.18	-0.10	0.31	0.24	-0.08	0.33	0.12	-0.16
TP_5	0.32	0.19	-0.03	0.30	0.18	-0.20	0.37	0.19	-0.12	0.39	0.14	-0.12
TP_6	0.19	0.21	-0.02	0.10	0.22	-0.23	0.23	0.17	-0.04	0.25	0.13	-0.05
TP_7	0.34	0.18	0.04	0.38	0.19	-0.25	0.35	0.22	-0.09	0.37	0.20	-0.04
TP_8	0.26	0.17	-0.04	0.30	0.22	-0.17	0.26	0.22	0.01	0.35	0.16	-0.02
TP_9	0.34	0.19	0.00	0.36	0.15	-0.16	0.30	0.13	-0.08	0.40	0.13	-0.08
TP_10	0.35	0.23	-0.05	0.33	0.20	-0.28	0.34	0.23	-0.11	0.36	0.16	-0.15
TP_11	0.38	0.17	-0.04	0.37	0.16	-0.22	0.36	0.17	-0.13	0.38	0.14	-0.13
TP_12	0.26	0.14	0.01	0.33	0.13	-0.11	0.22	0.10	-0.08	0.32	0.10	-0.05
TP_13	0.39	0.23	0.00	0.43	0.20	-0.32	0.43	0.19	-0.12	0.43	0.13	-0.11
TP_14	0.24	0.19	0.02	0.27	0.18	-0.23	0.27	0.18	-0.11	0.35	0.14	-0.08
TP_15	0.31	0.21	-0.06	0.30	0.23	-0.11	0.34	0.13	-0.13	0.32	0.13	-0.04
TP_16	0.31	0.20	0.00	0.32	0.21	-0.24	0.34	0.20	-0.08	0.34	0.15	-0.02
TP_17	0.24	0.20	0.01	0.26	0.17	-0.08	0.22	0.17	-0.01	0.31	0.18	-0.05
TP_18	0.37	0.20	0.01	0.42	0.19	-0.20	0.41	0.18	-0.06	0.41	0.13	-0.09
TP_19	0.34	0.17	-0.08	0.26	0.17	-0.13	0.26	0.16	-0.15	0.37	0.12	-0.11
TP_20	0.25	0.16	-0.04	0.30	0.14	-0.25	0.26	0.16	-0.11	0.30	0.12	-0.07
Mean Pre.acc	**0.32**	**0.19**	**-0.01**	**0.32**	**0.18**	**-0.19**	**0.32**	**0.17**	**-0.09**	**0.36**	**0.14**	**-0.08**

Table S4. 4: Prediction accuracies for optimized training population size of 800 for combined TPs (TP1 and TP2)

Train. Pop size 800	G-BLUP			Bayes-A			Bayes-B			Bayesian Lasso		
	M3CBSDs	M6CBSDs	CBSDRs	M3CBSDs	M6CBSDs	CBSDRs	M3CBSDs	M6CBSDs	CBSDRs	M3CBSDs	M6CBSDs	CBSDRs
TP_1	0.30	0.25	0.05	0.30	0.21	-0.13	0.27	0.21	-0.09	0.29	0.19	-0.01
TP_2	0.29	0.25	0.07	0.31	0.25	-0.09	0.27	0.25	-0.05	0.29	0.17	0.02
TP_3	0.31	0.26	0.06	0.34	0.26	-0.10	0.27	0.20	-0.05	0.33	0.15	0.00
TP_4	0.31	0.26	0.07	0.28	0.25	-0.14	0.30	0.25	0.03	0.30	0.17	0.03
TP_5	0.30	0.27	0.07	0.12	0.21	-0.13	0.28	0.27	-0.04	0.29	0.13	0.00
TP_6	0.30	0.27	0.06	0.33	0.26	-0.17	0.12	0.26	-0.05	0.29	0.18	-0.04
TP_7	0.30	0.27	0.07	0.35	0.22	-0.18	0.27	0.25	-0.01	0.27	0.17	0.00
TP_8	0.29	0.27	0.06	0.32	0.24	-0.15	0.30	0.28	-0.02	0.30	0.21	0.00
TP_9	0.30	0.26	0.07	0.33	0.27	-0.11	0.21	0.14	-0.05	0.28	0.16	0.00
TP_10	0.31	0.26	0.07	0.34	0.25	-0.20	0.38	0.26	-0.07	0.34	0.19	-0.05
TP_11	0.30	0.28	0.05	0.32	0.29	-0.13	0.35	0.33	0.01	0.32	0.16	-0.01
TP_12	0.30	0.26	0.07	0.34	0.27	-0.11	0.37	0.23	-0.03	0.34	0.18	0.01
TP_13	0.30	0.26	0.07	0.16	0.25	-0.12	0.34	0.13	-0.05	0.28	0.17	-0.03
TP_14	0.30	0.28	0.07	0.16	0.23	-0.20	0.33	0.21	-0.05	0.36	0.20	-0.03
TP_15	0.30	0.26	0.07	0.30	0.27	0.12	0.24	0.24	0.04	0.30	0.12	-0.01
TP_16	0.30	0.27	0.06	0.26	0.26	-0.18	0.30	0.30	-0.04	0.30	0.20	0.00
TP_17	0.30	0.26	0.06	0.31	0.25	-0.19	0.32	0.18	-0.10	0.37	0.16	-0.01
TP_18	0.30	0.27	0.06	0.34	0.26	-0.11	0.25	0.33	-0.05	0.30	0.19	-0.01
TP_19	0.29	0.25	0.05	0.29	0.25	-0.07	0.28	0.15	-0.07	0.30	0.13	-0.02
TP_20	0.28	0.25	0.06	0.26	0.23	-0.18	0.26	0.18	-0.05	0.29	0.11	-0.01
Mean Pre.acc	**0.31**	**0.26**	**0.06**	**0.29**	**0.25**	**-0.13**	**0.29**	**0.23**	**-0.04**	**0.31**	**0.17**	**-0.01**

Table S4. 5: Prediction accuracies for full set of training population (TP1 and TP2)

Train. Pop size 922	G-BLUP			Bayes-A			Bayes-B			Bayesian Lasso		
	M3CBSDs	M6CBSDs	CBSDRs	M3CBSDs	M6CBSDs	CBSDRs	M3CBSDs	M6CBSDs	CBSDRs	M3CBSDs	M6CBSDs	CBSDRs
Pre.acc	0.30	0.25	0.05	0.24	0.21	-0.11	0.30	0.26	-0.09	0.31	0.15	-0.04

Table S4. 6: Comparing prediction accuracies for optimized and random subset of training population of size 200

	Optimized Subset of TP			Random sub Set of TP		
Train. Pop size 200	**CBSD3s**	**CBSD6s**	**CBSDRs**	**CBSD3s**	**CBSD6s**	**CBSDRs**
TP_1	0.18	0.33	-0.42	0.30	0.01	-0.02
TP_2	0.27	0.38	-0.33	0.06	0.00	0.06
TP_3	0.09	0.34	0.08	0.40	0.15	-0.03
TP_4	0.39	0.29	0.33	0.25	0.06	-0.05
TP_5	0.40	0.25	0.16	0.36	-0.04	-0.04
TP_6	0.36	0.27	0.16	0.16	-0.03	-0.08
TP_7	0.28	0.29	0.30	0.50	0.04	-0.03
TP_8	0.29	0.20	0.05	0.15	-0.23	0.01
TP_9	0.26	0.36	-0.17	-0.18	-0.06	-0.12
TP_10	0.38	0.37	-0.17	0.07	-0.07	0.07
TP_11	0.26	0.29	-0.11	-0.07	0.07	0.11
TP_12	0.20	0.31	-0.12	-0.11	0.01	-0.05
TP_13	0.15	0.14	-0.16	0.06	-0.17	0.00
TP_14	0.22	0.22	-0.13	0.34	-0.04	-0.02
TP_15	0.28	0.31	-0.15	-0.13	0.03	-0.09
TP_16	0.25	0.17	0.18	-0.12	0.01	-0.09
TP_17	0.11	0.19	0.02	-0.07	-0.19	-0.12
TP_18	0.25	0.23	-0.18	-0.13	0.06	-0.09
TP_19	0.27	0.21	-0.19	0.23	0.09	-0.09
TP_20	0.38	0.27	0.12	0.21	0.11	0.02
Mean Pre.acc	**0.27**	**0.28**	**-0.04**	**0.11**	**-0.01**	**-0.03**

Table S4. 7: Comparing prediction accuracies for optimized and random subset of training population of size 400

Train. Pop size 400	Optimized Subset of TP			Random sub Set of TP		
	CBSD3s	CBSD6s	CBSDRs	CBSD3s	CBSD6s	CBSDRs
TP_1	0.40	0.18	0.01	0.17	-0.04	-0.03
TP_2	0.28	0.19	0.06	0.05	0.05	0.03
TP_3	0.32	0.16	0.01	0.15	0.14	-0.02
TP_4	0.34	0.21	-0.05	0.02	0.00	0.04
TP_5	0.32	0.19	-0.03	0.07	0.06	0.17
TP_6	0.19	0.21	-0.02	-0.02	0.16	-0.21
TP_7	0.34	0.18	0.04	0.19	-0.01	0.22
TP_8	0.26	0.17	-0.04	0.29	0.01	-0.01
TP_9	0.34	0.19	0.00	-0.10	0.10	-0.21
TP_10	0.35	0.23	-0.05	0.26	0.12	-0.12
TP_11	0.38	0.17	-0.04	0.38	0.11	0.07
TP_12	0.26	0.14	0.01	0.08	-0.08	-0.03
TP_13	0.39	0.23	0.00	-0.04	-0.12	0.11
TP_14	0.24	0.19	0.02	0.24	0.20	-0.02
TP_15	0.31	0.21	-0.06	0.17	0.09	0.04
TP_16	0.31	0.20	0.00	0.02	-0.05	0.08
TP_17	0.24	0.20	0.01	0.02	0.01	-0.02
TP_18	0.37	0.20	0.01	0.03	0.07	-0.28
TP_19	0.34	0.17	-0.08	0.23	0.01	0.08
TP_20	0.25	0.16	-0.04	-0.14	-0.11	0.01
Mean Pre.acc	**0.32**	**0.19**	**-0.01**	**0.10**	**0.04**	**-0.01**

Table S4. 8: Prediction accuracies for single and multi-kernel G-BLUP models for optimized training population size of 200 clones, where K_1, K_2 and K_3 represent single kernel, two kernels, and three kernels G-BLUP models respectively.

Training set 200	CBSD3s			CBSD6s			CBSDRs		
	K_1	K_2	K_3	K_1	K_2	K_3	K_1	K_2	K_3
TP1	0.18	0.27	0.30	0.34	0.44	0.45	-0.02	-0.01	0.09
TP2	0.27	0.29	0.32	0.39	0.46	0.42	0.06	0.07	-0.02
TP3	0.09	0.47	0.42	0.35	0.19	0.47	-0.03	-0.07	-0.06
TP4	0.38	0.42	0.43	0.29	0.45	0.38	-0.05	-0.06	-0.09
TP5	0.40	0.45	0.46	0.26	0.34	0.36	-0.04	-0.07	-0.15
TP6	0.35	0.31	0.33	0.29	0.34	0.41	-0.08	-0.21	-0.02
TP7	0.37	0.32	0.27	0.20	0.40	0.29	-0.03	-0.01	-0.03
TP8	0.29	0.30	0.35	0.37	0.30	0.50	0.01	-0.02	-0.12
TP9	0.26	0.35	0.26	0.37	0.49	0.47	-0.12	0.11	-0.04
TP10	0.37	0.30	0.21	0.30	0.47	0.43	0.07	-0.06	0.13
TP11	0.26	0.23	0.25	0.31	0.42	0.44	0.11	0.12	-0.05
TP12	0.20	0.19	0.22	0.15	0.27	0.28	-0.05	-0.02	-0.03
TP13	0.15	0.22	0.23	0.23	0.32	0.28	0.00	0.00	0.02
TP14	0.22	0.31	0.32	0.31	0.43	0.41	-0.02	-0.12	-0.11
TP15	0.28	0.27	0.22	0.17	0.17	0.17	-0.09	-0.15	-0.11
TP16	0.24	0.43	0.39	0.20	0.14	0.30	-0.09	-0.13	-0.12
TP17	0.11	0.34	0.34	0.23	0.33	0.26	-0.12	-0.15	-0.15
TP18	0.25	0.29	0.41	0.21	0.50	0.30	-0.09	-0.13	-0.12
TP19	0.26	0.20	0.26	0.38	0.25	0.33	-0.09	-0.13	-0.12
TP20	0.27	0.25	0.30	0.22	0.31	0.51	0.02	-0.01	0.24
Mean Pred	**0.27**	**0.31**	**0.32**	**0.28**	**0.35**	**0.37**	**-0.03**	**-0.05**	**-0.04**

Table S4. 9: Prediction accuracies for single and multi-kernel G-BLUP models for optimized training population of size 400, K_1, K_2 and K_3 represent single kernel, two kernels, and three kernels G-BLUP models respectively.

Training set 400	**CBSD3s**			**CBSD6s**			**CBSDRs**		
	K_1	K_2	K_3	K_1	K_2	K_3	K_1	K_2	K_3
TP1	0.40	0.42	0.41	0.17	0.24	0.26	0.00	0.01	0.02
TP2	0.28	0.24	0.23	0.19	0.34	0.26	0.06	-0.12	-0.10
TP3	0.32	0.39	0.34	0.16	0.23	0.25	0.01	0.00	0.01
TP4	0.34	0.31	0.34	0.21	0.27	0.29	-0.01	-0.09	-0.04
TP5	0.32	0.27	0.29	0.19	0.25	0.27	-0.05	-0.05	-0.03
TP6	0.19	0.43	0.43	0.21	0.30	0.30	-0.03	-0.04	-0.03
TP7	0.34	0.31	0.37	0.17	0.23	0.26	-0.02	0.03	0.04
TP8	0.26	0.24	0.28	0.17	0.27	0.28	0.04	-0.07	-0.08
TP9	0.34	0.29	0.36	0.19	-0.24	0.29	-0.04	-0.01	0.00
TP10	0.35	0.32	0.36	0.23	0.28	0.30	0.00	-0.05	-0.03
TP11	0.38	0.28	0.36	0.24	0.23	0.27	-0.05	-0.07	-0.04
TP12	0.26	0.27	0.28	0.14	0.20	0.23	-0.04	0.00	0.01
TP13	0.39	0.34	0.42	0.22	0.29	0.31	0.01	0.00	0.01
TP14	0.24	0.34	0.24	0.19	0.25	0.31	0.00	0.01	0.01
TP15	0.31	0.32	0.33	0.21	0.28	0.29	0.02	-0.01	-0.05
TP16	0.31	0.29	0.32	0.20	0.26	0.28	-0.06	0.00	0.01
TP17	0.24	0.29	0.28	0.20	0.23	0.29	0.00	0.02	0.04
TP18	0.37	0.39	0.40	0.20	0.24	0.26	0.01	0.00	0.01
TP19	0.34	0.29	0.30	0.17	0.20	0.27	0.01	-0.08	-0.06
TP20	0.25	0.34	0.49	0.16	0.25	0.23	-0.08	-0.01	-0.01
Mean Pred.	**0.32**	**0.32**	**0.34**	**0.19**	**0.23**	**0.27**	**-0.01**	**-0.03**	**-0.02**

Table S4 10: Five-fold cross validation, replicated 10 times for IITA clones using G-BLUP model.

Replications	CBSD3s	CBSD6s	CBSDRs
1	0.39	0.23	0.08
2	0.45	0.24	0.04
3	0.41	0.42	0.22
4	0.37	0.21	0.20
5	0.46	0.34	-0.03
6	0.37	-0.29	0.13
7	0.48	0.27	0.07
8	0.45	0.27	-0.20
9	0.36	0.22	0.13
10	0.33	0.14	0.18
Mean Pred.	**0.40**	**0.21**	**0.08**

Table S4. 11: Prediction accuracies of CBSD-traits for single and multi-kernel G-BLUP models under high density, whole genome sequence imputed markers (WGS) and low density genotyping-by-sequencing markers (GBS) markers for optimized training population size of 200

	Single Kernel model (K1)		Multi-Kernel (K2)		Multi-Kernel (K3)	
Traits	GBS markers	WGS Markers	GBS markers	WGS Markers	GBS markers	WGS Markers
CBSD3s	0.27	0.35	0.31	0.26	0.32	0.32
CBSD6s	0.28	0.15	0.35	0.24	0.37	0.22
CBSDRs	-0.03	0.18	-0.05	0.19	-0.04	0.14

Table S4 12: Prediction accuracies for CBSD related traits for single and multi-kernel G-BLUP models under high density, whole genome sequence imputed markers (WGS) and low density genotyping-by sequencing markers (GBS) markers for optimized training population size of 400

	Single Kernel model (K1)		Multi-Kernel (K2)		Multi-Kernel (K3)	
Traits	GBS markers	WGS Markers	GBS markers	WGS Markers	GBS markers	WGS Markers
CBSD3s	0.32	0.39	0.34	0.41	0.44	0.32
CBSD6s	0.19	0.15	0.23	0.16	0.20	0.27
CBSDRs	-0.01	0.16	-0.03	0.15	0.17	-0.02

Table S4. 13: Variance component and heritability estimates for TP1 and TP2.

Datasets	**TP1**				**TP2**			
Sources Variations	**CBSD3s**	**CBSD6s**	**CBSDRs**		**Sources Variations**	**CBSD3s**	**CBSD6s**	**CBSDRs**
Rep/Loc	0.127	0.013	0.017		Block/Loc	0.017	0.059	0.009
Clones	0.132	0.228	0.453		Clones	0.173	0.213	0.318
Clones x Loc	0.025	0.056	0.42		Clones x Loc	0.008	0.119	0.096
Residual	0.34	0.446	0.64		Residual	0.385	0.529	0.471
H2	**0.28**	**0.34**	**0.42**		**H2**	**0.31**	**0.29**	**0.40**

References

Akdemir, D., and U.G. Okeke. 2015. EMMREML: Fitting Mixed Models with known covariance structures,R Repository CRAN.

Akdemir, D., J.I. Sanchez, and J.L. Jannink. 2015. Optimization of genomic selection training populations with a genetic algorithm. Genet. Sel. Evol. 47: 1–10.

Alicai, T., J. Ndunguru, P. Sseruwagi, F. Tairo, G. Okao-Okuja, R. Nanvubya, L. Kiiza, L. Kubatko, M. a. Kehoe, and L.M. Boykin. 2016. Cassava brown streak virus has a rapidly evolving genome: implications for virus speciation, variability, diagnosis and host resistance. Sci. Rep. 6(October): 36164.

Alicai, T., C.A. Omongo, M.N. Maruthi, R.J. Hillocks, Y. Baguma, R. Kawuki, A. Bua, G.W. Otim-Nape, and J. Colvin. 2007. Re-emergence of Cassava Brown Streak Disease in Uganda. Plant Dis. 91(1): 24–29.

Beyene, G., R.D. Chauhan, M. Ilyas, H. Wagaba, C.M. Fauquet, D. Miano, T. Alicai, and N.J. Taylor. 2017. A Virus-Derived Stacked RNAi Construct Confers Robust Resistance to Cassava Brown Streak Disease. Front. Plant Sci. 7: 1–12.

Browning, B.L., and S.R. Browning. 2016. Genotype Imputation with Millions of Reference Samples. Am. J. Hum. Genet. 98(1): 116–126.

Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, and S.E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6: 1–10.

Endelman, J.B. 2011. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. Plant Genome J. 4: 250.

Garrick, D.J., J.F. Taylor, and R.L. Fernando. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet. Sel. Evol. 41: 55.

Glaubitz, J.C., T.M. Casstevens, F. Lu, J. Harriman, R.J. Elshire, Q. Sun, and E.S. Buckler. 2014. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS One 9(2): 90346.

Goiffon, M.D. 2016. Optimal population value selection: A population-based selection strategy for genomic selection. : 66-71.

Haffliger. 2016. Genomic predictions including known QTL for reproduction traits in swine. Master's thesis submitted to Norwegian University of Life Science.

Hayes, B.J., P.J. Bowman, A.J. Chamberlain, and M.E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: progress and challenges. J. Dairy Sci. 92: 433–443.

Heffner, E.L., J.-L. Jannink, and M.E. Sorrells. 2011. Genomic Selection Accuracy using Multifamily Prediction Models in a Wheat Breeding Program. Plant Genome 4: 65.

Heslot, N., H.-P. Yang, M.E. Sorrells, and J.-L. Jannink. 2012. Genomic Selection in Plant Breeding: A Comparison of Models. Crop Sci. 52: 146.

Hillocks, R. 2004. Research protocols for cassava brown streak virus. Fao.Org: 1–7.

Hillocks, R.J., and D.L. Jennings. 2003. Cassava brown streak disease: A review of present knowledge and research needs. Int. J. Pest Manag. 49: 225–234.

Hillocks, R., M. Maruthi, H. Kulembeka, S. Jeremiah, F. Alacho, E. Masinde, J. Ogendo, P. Arama, R. Mulwa, G. Mkamilo, B. Kimata, D. Mwakanyamale, A. Mhone, and I. Benesi. 2016. Disparity between Leaf and Root Symptoms and Crop Losses Associated with Cassava Brown Streak Disease in Four Countries in Eastern Africa. J. Phytopathol. 164: 86–93.

Hillocks, R.J., and J.M. Thresh. 2000. Cassava Mosaic and Cassava Brown Streak Virus Diseases in Africa : Root 7: 1–8.

Hillocks, R.J., J.M. Thresh, J. Tomas, M. Botao, R. Macia, and R. Zavier. 2002. Cassava brown streak disease in northern Mozambique. Int. J. Pest Manag. 48: 178–181.

Isidro, J., J.L. Jannink, D. Akdemir, J. Poland, N. Heslot, and M.E. Sorrells. 2015. Training set optimization under population structure in genomic selection. Theor. Appl. Genet. 128: 145–158.

Jannink, J.-L., A.J. Lorenz, and H. Iwata. 2010. Genomic selection in plant breeding: from theory to practice. Brief. Funct. Genomics 9: 166–177.

Kawuki, R.S., T. Kaweesi, W. Esuma, A. Pariyo, I.S. Kayondo, A. Ozimati, V. Kyaligonza, A. Abaca, J. Orone, R. Tumuhimbise, E. Nuwamanya, P. Abidrabo, T. Amuge, E. Ogwok, G. Okao, H. Wagaba, G. Adiga, T. Alicai, C. Omongo, A. Bua, M. Ferguson, E. Kanju, and Y. Baguma. 2016. Eleven years of breeding efforts to combat cassava brown streak disease. Breed. Sci. 66: 560–571.

Kayondo, S.I., D.P. Del Carpio, R. Lozano, A. Ozimati, M. Wolfe, Y. Baguma, V. Gracen, S. Offei, M. Ferguson, R. Kawuki, and J.L. Jannink. 2018. Genome-wide association mapping and genomic prediction for CBSD resistance in *Manihot esculenta*. Sci. Rep. 8: 1–11.

Legg, J.P., S.C. Jeremiah, H.M. Obiero, M.N. Maruthi, I. Ndyetabula, G. Okao-Okuja, H. Bouwmeester, S. Bigirimana, W. Tata-Hangy, G. Gashaka, G. Mkamilo, T. Alicai, and P. Lava Kumar. 2011. Comparing the regional epidemiology of the cassava mosaic and cassava brown streak virus pandemics in Africa. Virus Res. 159: 161–170.

Legg, J., E.A. Somado, I. Barker, L. Beach, H. Ceballos, W. Cuellar, J. Lorenzen, J. Lynam, M. Mcmahon, G. Maruthi, D. Miano, K. Mtunda, P. Natwuruhunga, E. Okogbenin, P. Pezo, E. Terry, G. Thiele, M. Thresh, J. Wadsworth, S. Walsh, S. Winter, J. Tohme, and C. Fauquet. 2014. A global alliance declaring war on cassava viruses in Africa. : 231–248.

Lorenz, A.J., S. Chao, F.G. Asoro, E.L. Heffner, T. Hayashi, H. Iwata, K.P. Smith, M.E. Sorrells, and J.L. Jannink. 2011. Genomic Selection in Plant Breeding. Knowledge and Prospects.

Lozano, R., D.P. Del Carpio, S.I. Kayondo, T. Amuge, O.A. Adebo, M. Ferguson, and J.-L. Jannink. 2017. Leveraging transcriptomics data for genomic prediction models in cassava. bioRxiv: 208181.

Lozano, R., M.T. Hamblin, S. Prochnik, and J.L. Jannink. 2015. Identification and distribution of the NBS-LRR gene family in the Cassava genome. BMC Genomics 16: 1–14.

Maruthi, M.N., R.J. Hillocks, K. Mtunda, M.D. Raya, M. Muhanna, H. Kiozia, A.R. Rekha, J. Colvin, and J.M. Thresh. 2005. Transmission of Cassava brown streak virus by *Bemisia tabaci* (Gennadius). J. Phytopathol. 153: 307–312.

Masona, M. V., N.J. Taylor, A.I. Robertson, and C.M. Fauquet. 2001. Transferring a cassava (*Manihot esculenta Crantz*) genetic engineering capability to the African environment: Progress and prospects. Euphytica 120: 43–48.

Masumba, E.A., F. Kapinga, G. Mkamilo, K. Salum, H. Kulembeka, S. Rounsley, J. V. Bredeson, J.B. Lyons, D.S. Rokhsar, E. Kanju, M.S. Katari, A.A. Myburg, N.A. van der Merwe, and M.E. Ferguson. 2017. QTL associated with resistance to cassava brown streak and cassava mosaic diseases in a bi-parental cross of two Tanzanian farmer varieties, Namikonga and Albert. Theor. Appl. Genet. 130: 2069–2090.

Mbewe, W., F. Tairo, P. Sseruwagi, J. Ndunguru, S. Duffy, S. Mukasa, I. Benesi, S. Sheat, M. Koerbler, and S. Winter. 2017. Variability in P1 gene redefines phylogenetic relationships among cassava brown streak viruses. Virol. J. 14: 1–7.

Meuwissen, T. H. E. , Hayes, B. J., & Goddard, M.E. 2001. Prediction of total genetic value

using genome-wide dense markers maps. Genetics 157: 1819–1829.

Mohammed, I.U., M.M. Abarshi, B. Muli, R.J. Hillocks, and M.N. Maruthi. 2011. The symptom and genetic diversity of cassava brown streak viruses infecting cassava in east africa. Adv. Virol. 2012: 10.

Monger, W., S. Seal, S. Cotton, and Foster. 2001. Identification of different isolates of cassava brown streak virus and developement of a diagnostic test. Plant Pathol. 50: 181–194.

Mulimbi, W., X. Phemba, B. Assumani, P. Kasereka, S. Muyisa, H. Ugentho, R. Reeder, J.P. Legg, L. Laurenson, R. Weekes, and F.E.. Thom. 2012. First report of Ugandan cassava brown streak virus on cassava in Democratic Republic of Congo. New Dis. Reports 26: 11.

Mware, B., E. Ateka, and J. Songa. 2009. Transmission and distribution of cassava brown streak virus disease in cassava growing areas of Kenya. J. Appl.864–870.

Nielsen, N.H., A. Jahoor, J.D. Jensen, J. Orabi, F. Cericola, V. Edriss, and J. Jensen. 2016. Genomic prediction of seed quality traits using advanced barley breeding lines. PLoS One 11(10): 1–19.

Njoroge, M.K., D.L. Mutisya, D.W. Miano, and D.C. Kilalo. 2017. Whitefly species efficiency in transmitting cassava mosaic and brown streak virus diseases. Cogent Biol. 3: 4–11.

Nweke. 2004. New challenges in the Cassava tarinsformation in Nigeria and Ghana: Environment and Production Technology Division International Food Policy Research Institute. Food Policy 67: 1–118.

Nzuki, I., M.S. Katari, J. V. Bredeson, E. Masumba, F. Kapinga, K. Salum, G.S. Mkamilo, T. Shah, J.B. Lyons, D.S. Rokhsar, S. Rounsley, A.A. Myburg, and M.E. Ferguson. 2017. QTL Mapping for Pest and Disease Resistance in Cassava and Coincidence of Some QTL with Introgression Regions Derived from Manihot glaziovii. Front. Plant Sci. 8: 1–15.

Odipio, J., E. Ogwok, N.J. Taylor, M. Halsey, A. Bua, C.M. Fauquet, and T. Alicai. 2014. RNAi-derived field resistance to Cassava brown streak disease persists across the vegetative cropping cycle. GM Crops Food 5: 16–19.

Ogwok, E., J. Odipio, M. Halsey, E. Gaitán-Solís, A. Bua, N.J. Taylor, C.M. Fauquet, and T. Alicai. 2012. Transgenic RNA interference (RNAi)-derived field resistance to cassava brown streak disease. Mol. Plant Pathol. 13: 1019–1031.

Patil, B.L., J.P. Legg, E. Kanju, and C.M. Fauquet. 2015. Cassava brown streak disease: A threat to food security in Africa. J. Gen. Virol. 96: 956–968.

Peixoto, L., L. Bhering, and C. Cruz. 2016. Determination of the optimal number of markers and individuals in a training population necessary for maximum prediction accuracy in F 2 populations by using genomic selection models. Genet. Mol. Res. Mol. Res 15: 17–29.

Prochnik, S., P.R. Marri, B. Desany, P.D. Rabinowicz, C. Kodira, M. Mohiuddin, F. Rodriguez, C. Fauquet, J. Tohme, T. Harkins, D.S. Rokhsar, and S. Rounsley. 2012. The Cassava Genome: Current Progress, Future Directions. Trop. Plant Biol. 5: 88–94.

R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. R Found. Stat. Comput. Vienna, Austria (ISBN 3-900051-07-0): 900051.

Rabbi, I., M. Hamblin, M. Gedil, P. Kulakow, M. Ferguson, A.S. Ikpan, D. Ly, and J.L. Jannink. 2014. Genetic mapping using genotyping-by-sequencing in the clonally propagated cassava. Crop Sci. 54: 1384–1396.

Ramu, P., W. Esuma, R. Kawuki, I.Y. Rabbi, C. Egesi, J. V. Bredeson, R.S. Bart, J. Verma, E.S. Buckler, and F. Lu. 2017. Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. Nat. Genet. 49: 959–963.

Rincent, R., D. Laloë, S. Nicolas, T. Altmann, D. Brunel, P. Revilla, V.M. Rodríguez, J. Moreno-Gonzalez, A. Melchinger, E. Bauer, C.C. Schoen, N. Meyer, C. Giauffret, C. Bauland, P. Jamin, J. Laborde, H. Monod, P. Flament, A. Charcosset, and L. Moreau. 2012. Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: Comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). Genetics 192: 715–728.

Roorkiwal, M., A. Rathore, R.R. Das, M.K. Singh, A. Jain, S. Srinivasan, P.M. Gaur, B. Chellapilla, S. Tripathi, Y. Li, J.M. Hickey, A. Lorenz, T. Sutton, J. Crossa, J.-L. Jannink, and R.K. Varshney. 2016. Genome-Enabled Prediction Models for Yield Related Traits in Chickpea. Front. Plant Sci. 7: 1666.

De Roos, A.P.W., B.J. Hayes, and M.E. Goddard. 2009. Reliability of genomic predictions across multiple populations. Genetics 183: 1545–1553.

Storey, H.H., and R.F.W. Nichols. 1938. Studies of the mosaic disease of cassava. Ann. Appl. Biol. 25(1906): 790–806.

Su, G., O.F. Christensen, L. Janss, and M.S. Lund. 2014. Comparison of genomic predictions using genomic relationship matrices built with different weighting factors to account for locus-specific variances. J. Dairy Sci. 97: 6547–6559.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414–4423.

Wagaba, H., G. Beyene, J. Aleu, J. Odipio, G. Okao-Okuja, R.D. Chauhan, T. Munga, H. Obiero, M.E. Halsey, M. Ilyas, P. Raymond, A. Bua, N.J. Taylor, D. Miano, and T. Alicai. 2017. Field Level RNAi-Mediated Resistance to Cassava Brown Streak Disease across Multiple Cropping Cycles and Diverse East African Agro-Ecological Locations. Front. Plant Sci. 7.

Wang, Q., Y. Yu, J. Yuan, X. Zhang, H. Huang, F. Li, and J. Xiang. 2017. Effects of marker density and population structure on the genomic prediction accuracy for growth trait in Pacific white shrimp Litopenaeus vannamei. BMC Genet. 18: 1–9.

Winter, S., M. Koerbler, B. Stein, A. Pietruszka, M. Paape, and A. Butgereitt. 2010. Analysis of cassava brown streak viruses reveals the presence of distinct virus species causing cassava brown streak disease in East Africa. J. Gen. Virol. 91: 1365–1372.

Wolfe, M.D., D.P. Del Carpio, O. Alabi, L.C. Ezenwaka, U.N. Ikeogu, I.S. Kayondo, R. Lozano, U.G. Okeke, A.A. Ozimati, E. Williams, C. Egesi, R.S. Kawuki, P. Kulakow, I.Y. Rabbi, and J.-L. Jannink. 2017. Prospects for genomic gelection in cassava breeding. Plant Genome 10: 1–19.

Wolfe, M.D., I.Y. Rabbi, C. Egesi, M. Hamblin, R. Kawuki, P. Kulakow, R. Lozano, D.P. Del Carpio, P. Ramu, and J.-L. Jannink. 2016. Genome-Wide Association and Prediction Reveals Genetic Architecture of Cassava Mosaic Disease Resistance and Prospects for Rapid Genetic Improvement. Plant Genome 9: 342–356.

Yu, X., X. Li, T. Guo, C. Zhu, Y. Wu, S.E. Mitchell, K.L. Roozeboom, D. Wang, M.L. Wang, G.A. Pederson, T.T. Tesso, P.S. Schnable, R. Bernardo, and J. Yu. 2016. Genomic prediction contributing to a promising global strategy to turbocharge gene banks. Nat. Plants 2: 1–7.

CHAPTER 5

GENERAL CONCLUSIONS

Genomic selection (GS) has revolutionized animal breeding and is gaining importance in planting breeding. One major advantage of GS over traditional phenotypic selection is that GS has the ability to shorten the breeding cycle. Initially, marker-assisted selection (MAS) was embraced in animal and plant breeding to shorten the breeding cycles; however, MAS only works well for qualitative traits. A number of important traits in crops species such as yield and quantitative disease resistance are controlled by many genes with small effects. Therefore, GS, which uses dense genome-wide markers, potentially captures the small effects. National Crops Resources Research Institute (NaCRRI), Uganda is piloting GS for cassava breeding to shorten the long breeding cycle (8-10 years). However, a number of questions remained initially unanswered in implementing GS. For example, what impact GS would have on the overall levels of genetic diversity and inbreeding in cassava due to this accelerated breeding cycle. We investigated these questions in chapter two of the thesis. We found that genetic diversity lost was not significant, and that less inbreeding occurred from initial cycle of genomic selection (C_0) to cycle one (C_1). Based on our investigation, GS is not expected to cause rapid inbreeding and loss of genetic diversity, as cassava breeding populations are advanced from one cycle of genomic selection to the next. Additionally, we assessed the phenotypic and genetic correlations between traits measured at seedling and clonal evaluations to answer the question of whether seedling data can inform clonal performance. For CBSD, we found strong genetic correlations between measures on seedling stage and later clonal evaluation of the genotypes. The application of the knowledge of seedling-clonal relationship ranges from using it to cull seedlings to direct use of the seedling data for building genomic prediction models. This information of high relationship between seedling and clonal

performances for cassava brown streak disease root severity (CBSDRs) has encouraged sharing of botanical seeds between the East African and West Africa breeding programs for pre-emptive breeding of CBSD resistance in Western germplasm. The information on W. African progeny evaluation at NaCRRI, a hot spot for CBSD will be used to assess the breeding values of West Africa progenitors, which progenitors can subsequently be used as parental stock in breeding for CBSD resistance in West Africa.

In chapter three of this thesis, we leveraged genomic and environmental data to further understand trait performance, considering G x E in a GS framework. Reasonable prediction accuracies were observed for three (CBSD3s), six (CBSD6s) and twelve months, i.e., at harvest, (CBSDRs) for cassava brown streak disease, harvest index (HI) and dry matter content (DMC) across genomic prediction models, used to address the different prediction challenges in real cassava breeding program such as; predicting the performance of newly generated seedlings (crosses), unobserved genotypes in unobserved environments as well as predicting for unobserved environments, hence cutting the initial cost of field evaluations. The most interesting result was that similar prediction accuracies were observed for the five traits indicated above in CV1 (prediction for unobserved genotypes) and CV2 prediction strategies (predicting performance of unobserved genotypes in never-evaluated environments, which is known to be the most complex prediction problem). The findings support the need to continue data collection of environmental variables, especially for prediction problem of unobserved genotypes evaluated in unobserved environments (CV2), where predictions are made via information sharing using the environmental covariates. Last but not least, for future G x E research in cassava, we recommend the involvement of multi-disciplinary scientists, especially the plant physiologists in order to pin-point the biological relevance of the environmental covariates measured. The involvement of multi-disciplinary scientists will provide better

decision supporting evidence in choosing the most appropriate environmental covariates to include in the G x E models for key cassava traits.

In chapter four of the thesis, our aim was to design a pre-emptive breeding strategy for CBSD, which was initially an endemic disease to the coastal region of east Africa. In the last two decades, however, CBSD has rapidly spread to cover the entire eastern, and parts of southern and central Africa. Concerns are emerging that CBSD could reach W. Africa, especially Nigeria the world's biggest cassava producer. Using genomic and phenotypic data generated at NaCRRI as a training set, we predicted CBSD in West Africa clones. We observed moderate prediction accuracy for CBSD foliar symptom; however, the prediction accuracy for CBSDR necrosis was generally low. In this situation, we recommended the initial efforts to pre-emptively breed for CBSD in West Africa to focus on testing progeny of germplasm of W. African origin in Eastern Africa similar to recommendations made in chapter two, and later use the progeny evaluation data to train CBSD prediction models for Western Africa clones. This initiative is already taking place based on the recommendation from this study. Recently, the NaCRRI cassava breeding program received over 5,000 seedlings from the Nigerian national cassava breeding program at the Nigerian National Root Crop Research Institute, which are being evaluated for CBSD resistance. The information from the progeny will be used to assess the breeding values of the progenitors of the progenies being evaluated. Furthermore, these data will be used to train GS for predicting CBSD in W. African clones. Hopefully higher prediction accuracies than in the present study can be achieved for CBSD resistance in W. African germplasm, which would strengthen the potential value of the proposed pre-emptive breeding strategy.